Copyright
by
Kalpana Seshadrinathan
2008

The Dissertation Committee for Kalpana Seshadrinathan
certifies that this is the approved version of the following dissertation:

Video Quality Assessment Based on Motion Models

Committee:

Alan C. Bovik, Supervisor

Wilson S. Geisler

Lawrence K. Cormack

Gustavo de Veciana

Sriram Vishwanath

Video Quality Assessment Based on Motion Models

by

Kalpana Seshadrinathan, B.Tech., M.S.

DISSERTATION

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT AUSTIN

August 2008

Dedicated to Appa, Amma, Archu and Vandu.

Acknowledgments

I did not want to write this section at the beginning when I started off writing my dissertation. Firstly, I was not really in a mood to write it when I thought about the monumental task of writing up the entire dissertation. Plus, I was not really sure I would be able to graduate as planned and it felt like writing the acknowledgments would jinx it. Most important of all, I used to dream of the day when I would have the leisure to write this and give it complete justice. Well, the day has finally arrived now that my dissertation draft has been turned in to the committee.

I am going to write this in chronological order of events that brought me to the U.S.A. to pursue my Ph.D. Well, Master's degree actually. I changed my mind down the lane.

First and foremost, I would like to thank my family for all their support and understanding. My parents for their constant encouragement, unwavering belief and confidence in me, for giving me everything I ever wanted - including my unmarried status while I studied. My sister, Archu, for always looking out for me and making it possible for me to come to UT. My sister, Vandu, for always taking care of me. My brother-in-law, Amitesh, for his constant support and being a fellow "techie" in the family. My cousin, Nandu anna, for all his encouragement and help with my graduate applications - I wouldn't be here if it were not for him. My cousins in the U.S. for giving me the emotional support that I needed when I first arrived all alone in the U.S.A.

I would like to thank two of my teachers from my high school days, who really challenged my intellectual ability and inspired me - K.S.R. and Sankaran Sir. They showed me the wonderful world of mathematics and I learnt about my own interests and abilities. I can say with full confidence that my educational journey since would not have been the same without them.

I would like to thank Prof. Baxter Womack for offering me a TA in my first year at UT and all the hugs since.

The biggest thanks of all goes to Prof. Al Bovik, my adviser. For taking me on as a student and filling the role of adviser in every sense of the word. I wish every student in this world could have an adviser as great as Dr. Bovik. He inspired me with his teaching, which was the reason I joined his lab. He continues to inspire and awe me with his razor-sharp intellect and insight. He is always looking out for his students and I would like to thank him for his endless patience, understanding, and consideration. Words fail me in describing what it means to me to have had the good fortune to be his student.

I would like to thank Dr. Bovik and Golda for all the wonderful times we had at conferences and at parties in their house. They are two of the most wonderful people, with the unique ability of never making you feel like an outsider. Dr. Bovik once even told me that I was family and I can't even begin to explain what that meant to me. I want to thank Dhivya and Mercy for some wonderful times and giving me the joy of being with kids, something I have always missed since I came to UT. I remember I was very nervous about how they would react to me when I first met them in Italy. I had always got along well with kids in India, but was not sure if American kids would like me. I learnt that kids are the same everywhere. I would also like to thank Golda's dad for his support and affection.

I would like to thank all the members of my dissertation committee. In particular, a big thanks to Dr. Geisler and Dr. de Veciana for writing letters of recommendation for me. Also, a special thanks to Prof. de Veciana for advice and taking the time to talk to me several times during my 6 years at UT. I would like to thank Dr. Cormack for his help with the subjective study. I would like to thank Prof. Vishwanath for agreeing to serve in my committee at the last moment. I would like to thank Prof. Garg for serving in my qualifying exam committee.

I would like to thank all the wonderful teachers I have had here at UT for sharing the wonderful gift of learning. I would like to thank Pierre Costa for several years of support and both Pierre and Djoko Astronoto for a great working relationship at AT&T Research. I would like to thank many others at UT for their support. Melanie Gulick, the friendliest face you see when you first come to UT and the friendliest face you will ever see the rest of your time at UT. Shirley Watson, Selina Keilani, Janet Preuss and Paul White for all the help with administrative matters and appointments. Mary Matejka and

Gabriel Hernandez for the IT support. Merydith Turner for everything else.

Finally, my friends. Anyone who has been a member of LIVE in the "Golden age of Umesh" cannot thank anyone else before him, most of all me. I would like to thank Umesh for all the wonderful times together playing ping pong, tennis, learning to swim, sampling Austin restaurants, watching UT football games in the magic year of the national championship and so much more. Not to mention always giving me feedback on posters and papers, helping me pack, bringing me food when I came back depressed after an India trip. Umesh is solely responsible for patiently teaching me what little hand-eye coordination I have today. I would like to thank him for being my rock.

I would like to thank Balaji for being my first real friend at UT and all the support, encouragement and disagreements since. He helped me grow into a more mature person in many ways and I would like to thank him for always being there for me. A big thanks for sharing Rowdy with me this last year, helping me get over my fear and learning to love Rowdy and other animals.

I would like to thank Abraham for all the support the past 10 years.

I would like to thank Farooq for eating all the food I cooked, his sense of humor and being a wonderful friend. Hamid for his loud, amusing, annoying yet endearing personality and always being a good friend. Sumohana for his support, some wonderful times together at AT&T, help anytime I ever asked him for anything and putting up with my tantrums. Mehul, for being a wonderful friend - always considerate, affectionate and looking out for me. I would like to thank Krupa and Pallavi for the great parties and movies. Joonsoo, for always helping me out when I was in a jam and his sense of humor that made the lab so much fun. Sina, for his unique, wonderfully outgoing personality that I wish everyone in this world could have and forgetting to invite me to movies. Liu, for all the research discussions that bordered on fighting, initiating us into Chinese movies, games and culture, his amazing intelligence and his great sense of humor. James - for being his funny and entertaining self, for giving us a glimpse of American life and that wonderful day on his boat. Choi, for all the marriage advice. Abtine, for his friendship. Ian, for his friendliness and excellent imitation of Abtine in the gym. Tom Arnow for his wacky sense of humor and interest in Indian culture. I also want to thank Hyung, Rajiv, Anush, William, Rana, Shalini, Raghu and Preethi for some great times at LIVE.

Video Quality Assessment Based on Motion Models

Publication No. ____________________

Kalpana Seshadrinathan, Ph.D.
The University of Texas at Austin, 2008

Supervisor: Alan C. Bovik

A large amount of digital visual data is being distributed and communicated globally and the question of video quality control becomes a central concern. Unlike many signal processing applications, the intended receiver of video signals is nearly always the human eye. Video quality assessment algorithms must attempt to assess *perceptual degradations* in videos. My dissertation focuses on full reference methods of image and video quality assessment, where the availability of a perfect or pristine reference image/video is assumed.

A large body of research on image quality assessment has focused on models of the human visual system. The premise behind such metrics is to process visual data by simulating the visual pathway of the eye-brain system. Recent approaches to image quality assessment, the structural similarity index and information theoretic models, avoid explicit modeling of visual mechanisms and use statistical properties derived from the images to formulate measurements of image quality. I show that the structure measurement in structural similarity is equivalent to contrast masking models that form a critical component of many vision based methods. I also show the equivalence of the structural and the information theoretic metrics under certain assumptions on the statistical distribution of the reference and distorted images.

Videos contain many artifacts that are specific to motion and are largely temporal. Motion information plays a key role in visual perception of video signals. I develop a general, spatio-spectrally localized multi-scale framework for evaluating dynamic video fidelity that integrates both spatial and temporal aspects of distortion assessment. Video quality is evaluated in space and time by evaluating motion quality along computed motion trajectories. Using this framework, I develop a full-reference video quality assessment algorithm known as the MOtion-based Video Integrity Evaluation index, or MOVIE index.

Lastly, and significantly, I conducted a large-scale subjective study on a database of videos distorted by present generation video processing and communication technology. The database contains 150 distorted videos obtained from 10 naturalistic reference videos and each video was evaluated by 38 human subjects in the study. I study the performance of leading, publicly available objective video quality assessment algorithms on this database.

Table of Contents

Acknowledgments **v**

Abstract **x**

List of Tables **xv**

List of Figures **xvi**

Chapter 1. Introduction **1**

- 1.1 Motivation . . . 1
- 1.2 Concepts in Quality Assessment . . . 2
- 1.3 Contributions . . . 4
  - 1.3.1 Understanding Image Quality Assessment Methods . . . 5
  - 1.3.2 Spatio-temporal Quality Assessment of Natural Videos . 6
  - 1.3.3 Subjective Quality Assessment of Natural Videos 8

Chapter 2. Background **11**

- 2.1 Image Quality Assessment . . . 11
  - 2.1.1 HVS-based metrics . . . 12
  - 2.1.2 Signal Fidelity based Methods . . . 15
- 2.2 Video Quality Assessment . . . 17
- 2.3 Conclusion . . . 21

Chapter 3. Unified Treatment of Full Reference Image Quality Assessment Algorithms **23**

- 3.1 Notation . . . 26
- 3.2 Structural Similarity Metrics . . . 27
  - 3.2.1 The SSIM Index . . . 27
  - 3.2.2 Probabilistic SSIM Index . . . 30

3.2.3 Relation to MSE 31
3.2.4 Relation to HVS Based Metrics 32
3.2.5 Discussion 38
3.3 The Information Theoretic Metrics 44
3.3.1 The IFC and VIF Indices 44
3.3.1.1 The Information Fidelity Criterion 44
3.3.1.2 The Visual Information Fidelity Criterion . . . 46
3.3.2 Relation of IFC to SSIM 48
3.3.2.1 Scalar Model 48
3.3.2.2 Vector Model 51
3.3.3 Relation of VIF to SSIM 55
3.3.3.1 Scalar Model 56
3.3.3.2 Vector Model 60
3.3.4 Discussion 61
3.4 Conclusion 66

Chapter 4. Spatio-temporal Quality Assessment of Natural Videos 68
4.1 Distortions in Digital Video 71
4.1.1 Spatial Artifacts 72
4.1.2 Temporal Artifacts 73
4.2 Spatio-temporal framework for video quality assessment 74
4.2.1 Linear Decomposition 74
4.2.2 Spatial MOVIE Index 78
4.2.3 Motion Estimation 84
4.2.4 Temporal MOVIE Index 85
4.2.5 Pooling Strategy 91
4.2.6 Implementation Details and Examples 95
4.3 Relation to Existing Models 99
4.3.1 Spatial Quality Computation 99
4.3.2 Temporal Quality Computation 100
4.4 Performance 102
4.5 Conclusion 105

Chapter 5. Study of subjective and objective quality assessment of video 107

5.1 Details of Subjective Study 110
  5.1.1 Source Sequences 110
  5.1.2 Test Sequences 113
    5.1.2.1 MPEG-2 compression 114
    5.1.2.2 H.264 compression 118
    5.1.2.3 Transmission over IP Networks 118
    5.1.2.4 Transmission over wireless networks 121
  5.1.3 Subjective Testing Design 122
  5.1.4 Subjective Testing Display 124
  5.1.5 Subjects and Training 125
5.2 Processing of Subjective Scores 126
5.3 Objective QA Models 128
5.4 Performance of Objective Models 130
5.5 Conclusions 134

Chapter 6. Conclusions and Future Work 135

6.1 Pooling Strategies 136
6.2 Scalable Quality Assessment 137
6.3 Reduced Reference Quality Assessment 138
6.4 Natural Scene Statistics 138
6.5 Blind Quality Assessment 139

Appendix 140

Appendix 1. Optical Flow Computation Via a New Multi-scale Approach 141

Bibliography 144

Vita 162

List of Tables

4.1 Comparison of the performance of VQA algorithms. 103
4.2 Performance of MOVIE on the VQEG database after omitting the animation sequences. 104

5.1 Comparison of the performance of VQA algorithms - SROCC 131
5.2 Comparison of the performance of VQA algorithms - LCC . . 131

List of Figures

1.1 Example showing (a) a reference image and (b) a test image that are available to a full reference IQA algorithm. 4

2.1 Block diagram of HVS-based quality metrics 12
2.2 Block diagram of SSIM quality assessment system 15
2.3 Block diagram of VIF quality assessment system 16
2.4 Block diagram of HVS-based video quality assessment system 18

3.1 Illustration of the need for the luminance and contrast comparison terms. 43
3.2 Plot of the VIF index as a function of $\sigma^2_{\mathbf{x}(\mathbf{i},k)}$ 57
3.3 Plot of $P_{\text{SSIM}}(C_3)$ as a function of $\sigma_{\mathbf{f}(\mathbf{i})}\sigma_{\mathbf{g}(\mathbf{i})}$ and C_3 60
3.4 Plot of $P_{\text{VIF}}(\kappa)$ as a function of $\sigma^2_{\mathbf{x}(\mathbf{i},k)}$ and κ 60

4.1 Geometry of the Gabor filterbank in the frequency domain. The figure shows iso-surface contours of all Gabor filters at the finest scale. The two horizontal axes denote the spatial frequency coordinates and the vertical axis denotes temporal frequency. . 79
4.2 A slice of the Gabor filter bank along the plane of zero temporal frequency. The x-axis denotes horizontal spatial frequency and the y-axis denotes vertical spatial frequency. 79
4.3 A slice of the Gabor filter bank along the plane of zero vertical spatial frequency. The x-axis denotes horizontal spatial frequency and the y-axis denotes temporal frequency. 80
4.4 A slice of the Gabor filters and the spectral plane shown in 2 dimensions. The horizontal axis denotes horizontal spatial frequencies and the vertical axis denotes temporal frequencies. Each circle represents a Gabor filter and the centers of each filter are also marked. The radius ρ of the single scale of Gabor filters and the the distance δ of the center frequency of one Gabor filter from the spectral plane are marked. 92

4.5 Illustration of the performance of the MOVIE index. Top left shows a frame from the reference video. Top right shows the corresponding frame from the distorted video. Bottom left shows a logarithmically compressed temporal quality map. Bottom right shows the spatial quality map. Notice that the spatial quality map responds to the blur in the test video. The temporal quality map responds to motion compensation mismatches surrounding the harp and the heads of the two people and distortions in the strings. 97

4.6 Illustration of the performance of the MOVIE index. Top left shows a frame from the reference video. Top right shows the corresponding frame from the distorted video. Bottom left shows a logarithmically compressed temporal quality map. Bottom right shows the spatial quality map. Notice that the spatial quality map responds to the blur in the test video. The temporal quality map responds to motion compensation mismatches surrounding the man, the oar and the ripples in the water. . . 98

4.7 Scatter plot of the subjective DMOS scores against MOVIE scores. The best fitting logistic function used for non-linear regression is also shown. (a) On all sequences in the VQEG database (b) After omitting the animated videos. 105

5.1 One frame from each of the reference video sequences used in the study. 112

5.2 One frame from each of the reference video sequences used in the study. 113

5.3 (a) Frame of the reference "riverbed" sequence in the LIVE database (b) Frame from one of the MPEG-2 compressed test video 115

5.4 (a) Frame from one of the H.264 compressed test video (b) Frame from test video distorted in a simulated IP network . . 116

5.5 Frame from test video distorted in a simulated wireless network 117

5.6 Scatter plots of objective VQA scores vs. DMOS for all videos in the LIVE database. Also shown is the best fitting logistic function. (a) PSNR (b) SSIM (c) Multi-scale SSIM (d) Speed SSIM 132

5.7 Scatter plots of objective VQA scores vs. DMOS for all videos in the LIVE database. Also shown is the best fitting logistic function. (a) VSNR (b) Spatial MOVIE (c) Temporal MOVIE (d) MOVIE 133

Chapter 1

Introduction

1.1 Motivation

The arrival of the personal computer and the Internet has ushered in a remarkable digital video revolution, the products of which pervade our daily lives in many ways. Digital video acquisition, communication, storage and display devices have advanced to an extraordinary degree of efficiency, leading to the rapid rise of popular applications such as Internet Video, Interactive Video on Demand (VoD), Wireless Video, HDTV, Digital Cinema and so on. With such a large amount of digital visual data being distributed and communicated globally, it is natural that the question of video quality control become a central concern. Unfortunately, rapid advances in video processing and communication have not been matched by similar progress in methods for video quality analysis. Unlike many signal processing applications, the intended receiver of video signals is nearly always the human eye. Image quality assessment (IQA) and video quality assessment (VQA) algorithms refer to automatic methods that attempt to predict *perceptual degradations* in images or videos. In other words, objective IQA/VQA algorithms quantify the quality of a given image/video *as seen by a human observer*.

The development of successful VQA algorithms will enormously further the cause of digital services to consumers, as well as researchers. Applications such as video-on-demand, wireless video services, digital cinema, and video conferencing with mobile devices will particularly benefit from this research, while new algorithms in biomedical imaging, astronomy, geophysics (and many others) will benefit from the availability of effective IQA/VQA algorithms as a means of testing algorithm results. The design of algorithms for image and video processing based on *perceptual criteria* is a nascent area of research that is driven by effective IQA/VQA models. Finally, IQA/VQA research promises to improve our basic understanding of human visual information processing, specifically with regard to mechanisms that contribute to quality perception.

1.2 Concepts in Quality Assessment

I begin by reviewing some of the essential concepts involved in IQA and VQA. Subjective judgment of quality must be regarded as the ultimate standard of performance by which IQA/VQA algorithms are assessed. Subjective quality is measured by displaying images or videos to human observers. The subject then indicates a quality score on a numerical or qualitative scale. To account for human variability and to assert statistical confidence, multiple subjects are required to view each image/video, and a Mean Opinion Score (MOS) is computed. While subjective IQA/VQA is the only completely reliable method, subjective studies are cumbersome, expensive, and more complex than they may seem. For example, statistical significance of the MOS must be guaranteed by using sufficiently large sample sizes; subject naivety must be imposed; the dataset of images/videos must be carefully calibrated; the display monitors should be calibrated; and so on. Subjective QA methods are impractical for nearly every application other than benchmarking automatic IQA/VQA algorithms.

There are three loosely agreed-upon categories of objective algorithms.

Full reference algorithms operate on distorted images while having a pristine, ideal reference image available for comparison. The vast majority of IQA and VQA algorithms fall into this category, because of the relative simplicity of making quality judgments relative to a standard. An example reference and distorted image are shown in Fig. 1.1.

Reduced reference algorithms operate without the use of a pristine reference, but do make use of additional (side) information along with the distorted image or video signal. Reduced reference algorithms may use features such as localized spatio-temporal activity information, edge locations extracted from an original reference, or embedded marker bits in the video stream as side information to estimate the distortion of the channel [1]. Other algorithms use knowledge that has been independently derived regarding the distortion process, such as foreknowledge of the nature of the distortion introduced by a compression algorithm, e.g., blocking, blurring, or ringing. Sometimes algorithms of this latter type are referred to as “blind”, but in my view, these should be categorized separately or as reduced reference algorithms.

(a) (b)

Figure 1.1: Example showing (a) a reference image and (b) a test image that are available to a full reference IQA algorithm.

No reference algorithms, also known as blind methods, attempt to assess the image/video quality without using any information other than the distorted signal. This process has proved daunting and there is very little substantive work on this topic. Yet, human beings can perform the task almost instantaneously, which suggests that there is hope in this direction, but in the long term.

I believe that much yet remains to be learned regarding full reference and reduced reference techniques, and especially regarding human visual perception of quality, before generic no reference algorithms become feasible. This dissertation deals exclusively with full reference methods of IQA and VQA.

1.3 Contributions

The following is an overview of the contributions presented in this dissertation.

1.3.1 Understanding Image Quality Assessment Methods

Diverse approaches to the full reference IQA problem have been proposed over the past three decades. A large class of methods, broadly classified as human visual system (HVS) based methods, build models of several stages of low level processing in the HVS and pass the reference and distorted images through these models to compute perceptual quality. The Structural SIMilarity (SSIM) index attempts to quantify the loss of structural distortions in images, arguing that humans equate loss of structural information with visual quality. Information theoretic methods, such as the Information Fidelity Criterion (IFC) and the Visual Information Fidelity (VIF) index, equate loss of visual quality with the amount of information that can be extracted by humans from the reference and distorted images using models of natural scene statistics.

I analyze the SSIM and information theoretic philosophies for full reference image quality evaluation in a general probabilistic framework to deepen our understanding of these indices and enable a unification of ideas derived from different first principles. I also explore the relationship between the SSIM index and models of contrast gain control used in HVS-based methods. I show that the structure term of the SSIM index is equivalent to certain models of contrast gain control in HVS-based methods and can hence account for contrast masking effects in human vision. I establish the equivalence of IFC and multi-scale SSIM models, assuming that the same linear decomposition is used in both models and that the statistical model assumed by the IFC is valid on the resulting scale-space representation of the images. I also study the relationship between the SSIM and VIF indices and reveal certain instabilities in the VIF formulation. My analysis shows a unifying link between certain IQA models based on signal statistics and IQA models based on visual processing of these signals. My analysis also reveals the strengths and weaknesses of different methods and it is my hope that future research into IQA techniques will benefit from this.

I describe the unified treatment of full reference of IQA algorithms in Chapter 3.

1.3.2 Spatio-temporal Quality Assessment of Natural Videos

VQA has traditionally been addressed using simple extensions of IQA methods to handle the temporal dimension. Although current full reference VQA algorithms incorporate features for measuring spatial distortions in video signals, very little effort has been spent on directly measuring temporal distortions or motion artifacts. While video signals do suffer from spatial distortions, they are also degraded by severe *temporal* artifacts such as ghosting, motion compensation mismatch, jitter, smearing, mosquito noise amongst numerous other types. Motion plays a very important role in visual perception of videos. Humans are very sensitive to motion, and considerable resources in the HVS are expended on computation of the velocity and direction of motion of image intensities from the time-varying images captured by the retina. People have the ability to execute smooth pursuit eye movements and visual attention is drawn to moving objects in a scene. Hence, visual perception of motion plays a very important role in video quality assessment. It is imperative that video quality indices account for both visual perception of motion and the deleterious effects of temporal artifacts, if objective evaluation of video quality is to accurately predict subjective judgment.

I develop a general framework for achieving spatio-spectrally localized multi-scale evaluation of dynamic video quality. In this framework, both spatial and temporal (and spatio-temporal) aspects of distortion assessment are accounted for. Video quality is evaluated not only in space and time, but also in space-time, by evaluating motion quality along computed motion trajectories. Using this framework, I develop a full reference VQA algorithm known as the MOtion-based Video Integrity Evaluation index, or MOVIE index. MOVIE integrates explicit motion information into the VQA process by tracking perceptually relevant distortions along motion trajectories, thus augmenting the measurement of spatial artifacts in videos. I validate the performance of MOVIE on a publicly available database of videos and demonstrate significant improvements using such an approach in matching visual perception.

I have sought to use principles derived from the analysis of IQA methods to improve the ability of MOVIE to capture spatial distortions in video. I also show how models used to track video quality along temporal trajectories in MOVIE relate to computational models of motion perception in the HVS. I believe that my work delivers the much-needed tool of motion modeling and temporal distortion modeling to the VQA community, to reach the ultimate goal of matching human perception.

I describe the framework for spatio-temporal evaluation of video quality and the MOVIE index in Chapter 4.

1.3.3 Subjective Quality Assessment of Natural Videos

I conducted a study to assess the subjective quality of videos. The study included 10 raw naturalistic reference videos and 150 distorted videos obtained from the references using four different real world distortion types. Each video was assessed by 38 human subjects using a single stimulus, continuous quality scoring procedure. The resulting database of videos is known as the Laboratory for Image and Video Engineering (LIVE) Video Quality Assessment Database.

Currently, the only publicly available subjective data that is widely used in the VQA community comes from the study conducted by the Video Quality Experts Group (VQEG) as part of its FR-TV Phase I project in 2000 [2]. This database has several limitations that I seek to address in my study. The videos in the LIVE VQA Database are all captured in progressive scan formats. The LIVE video quality database includes videos distorted by MPEG-2 and H.264 compression, as well as videos resulting from the transmission of H.264 packetized streams through error prone Internet Protocol (IP) and wireless communication channels. The LIVE database spans a wide range of quality - the low quality videos were designed to be of similar quality as videos typical of streaming video applications on the Internet (for example, Youtube). Great care was taken to ensure precise display of these psychophysical stimuli to human subjects in my study.

The goal of my study was to develop a database of videos that will challenge automatic VQA algorithms. I included *diverse* distortion types to test the ability of objective models to predict visual quality consistently across distortions. Compression systems such as MPEG-2 and H.264 produce fairly uniform distortions/quality in the video, both spatially and temporally. Network losses, however, cause *transient* distortions in the video, both spatially and temporally that appear as glitches in the video. My database is unique in this respect, since the VQEG Phase I database does not include such spatio-temporally localized distortion types.

Additionally, I adjusted the distortion strengths manually so that the videos obtained from each source and each distortion category spanned the same range of visual quality. This tests the ability of objective VQA models to predict visual quality across content and distortion types consistently.

I present an evaluation of the performance of leading, publicly available objective VQA algorithms on this database. The LIVE database provides a valuable tool to researchers in the VQA community for performance evaluation and advancement of current and future VQA algorithms.

I describe the LIVE VQA database and the performance evaluation of objective VQA algorithms on this database in Chapter 5.

I conclude this dissertation, with a discussion of avenues for future work in Chapter 6.

Chapter 2

Background

In this chapter, I review previous work on the full reference image quality assessment (IQA) and video quality assessment (VQA) problems. In Section 2.1, I review previous work on full reference IQA. In Section 2.2, I describe previous progress in the development of algorithms for VQA.

2.1 Image Quality Assessment

A large body of work has focused on using models of the human visual system (HVS) to develop quality metrics, broadly classified as HVS-based metrics. The basic idea behind these approaches is that the best way to predict the quality of an image, in the absence of any knowledge of the distortion process, is to attempt to "see" the image using a system similar to the HVS. I describe some well known HVS-based metrics in Section 2.1.1. I briefly describe some quality assessment models based on signal fidelity measures in Section 2.1.2. A review of quality assessment techniques for still images can be found in [3].

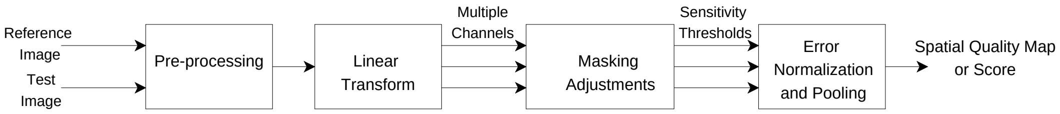


Figure 2.1: Block diagram of HVS-based quality metrics

2.1.1 HVS-based metrics

The earliest and most commonly used full reference quality metric is the Mean Squared Error (MSE) or equivalently, the Peak Signal to Noise Ratio (PSNR). This metric is popular due to its simplicity and mathematical tractability, although it is well known that it correlates poorly with visual quality [4]. PSNR is a simple mathematical measure and belongs to the class of metrics that do not incorporate any knowledge of the HVS. Other simple measures that weight different regions in frequency space differently have also been proposed as quality metrics and a discussion of these approaches can be found in [5]. These metrics can be thought of as a refinement to PSNR, since they incorporate modeling of some aspects of the HVS in designing the weights. In general, however, they have not been found to correlate well with visual quality across different images and varying types of distortions.

The premise behind HVS-based metrics is to process the visual data by simulating the visual pathway of the eye-brain system. As depicted in Fig. 2.1, HVS-based IQA systems typically begin by preprocessing the signal to correct for non-linearities, since lightness perception is a non-linear function of luminance. A filterbank decomposes reference and distorted or test signals into multiple spatial frequency- and orientation-tuned channels in an attempt to model similar processing by the cortical neurons. The *luminance masking*, *contrast masking* and *contrast sensitivity* features of the HVS are then modeled to account for perceptual error visibility. The response of the HVS to variations in luminance is a nonlinear function of the local mean luminance and this is commonly referred to as luminance masking. It is called masking because the variations in the distorted signal are masked by the base luminance of the reference image. Contrast masking refers to the reduction in visibility of one frequency component due to the presence of a stronger component of similar frequency or orientation in adjacent spatial locations. Additionally, the HVS has a bandpass characteristic and the frequency response is described by the Contrast Sensitivity Function (CSF). Baseline contrast sensitivity is the minimum amount of energy required to detect a particular channel component and can be computed for each channel using the CSF. The masking block uses the baseline contrast sensitivity, models for luminance and contrast masking and the image component in each channel to compute the sensitivity threshold for that channel. A space-varying threshold map is then created for each channel, which describes the sensitivity of each spatial location to errors in that particular channel. In the final stage, the error between the reference and test images in each channel is normalized by its corresponding sensitivity threshold and these normalized errors are known as Just Noticeable Differences (JND's). Finally, the JND values for all the channels at each spatial location are combined, usually using the Minkowski error metric, to generate a space-varying map of the image. This map predicts the probability that an observer will be able to detect any difference between the two images in local regions and can be combined suitably to generate a single number that represents the quality of the entire image, if desired.

This approach to IQA is intuitive and has met with considerable success. Different quality metrics proposed in the literature use different models for the blocks shown in Fig. 2.1. Popular approaches that followed the above paradigm include the pioneering work by Mannos and Sakrison [6], Lubin's laplacian-pyramid-based approach [7, 8] used in the Emmy award winning Sarnoff JNDMetrix technology [9], Daly's Visible Differences Predictor [10] using the Cortex Transform [11] and Teo and Heeger's steerable pyramid approach [12, 13].

Many of these models attempt to predict whether an observer will successfully discriminate between the reference and distorted images [7, 9, 10]. Discriminability, however, does not necessarily equate to visual quality, since different visible distortions have different *annoyance levels*. Further, most of these models are derived using stimuli such as sine waves and Gabor patches. The applicability of such models to natural images is questionable, in view of the highly non-linear nature of visual processing. Visual Signal to Noise Ratio (VSNR) is a method that attempts to ameliorate these effects [14]. Firstly, the computational models used in VSNR are derived based on psychophysical experiments conducted to quantify the visual detectability of distortions in natural images. Second, VSNR attempts to quantify the perceived contrast of supra-threshold distortions and the model is not restricted to the regime of

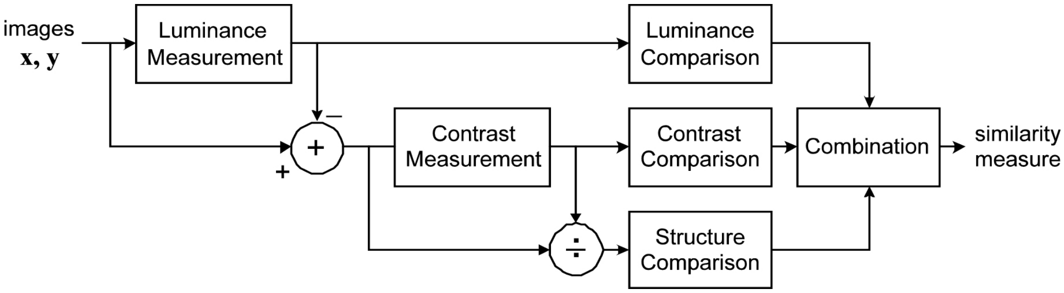


Figure 2.2: Block diagram of SSIM quality assessment system

threshold of visibility. Third, VSNR attempts to capture a mid-level property of the HVS known as global precedence, while most other models discussed here only consider low level processes in the visual system.

Different HVS-based approaches have had varying degrees of success. However, all of these methods suffer from certain drawbacks [15]. Although a lot is known about the early stages of the visual pathway, vision science has a long way to go before arriving at a clear understanding of the functioning of the entire HVS. Developing computational models of human vision is also an active research area and currently, much of the work on computational modeling is restricted to low level visual processing. A HVS-based quality metric can only be as good as the underlying model of human vision which is imperfect, to say the least, today [1, 16]. Further, HVS-based metrics generally require extensive calibration derived from human studies to determine the model parameters.

2.1.2 Signal Fidelity based Methods

The Structural SIMilarity (SSIM) approach to IQA assumes that the HVS has evolved to extract *structural information* from an image [15, 17].

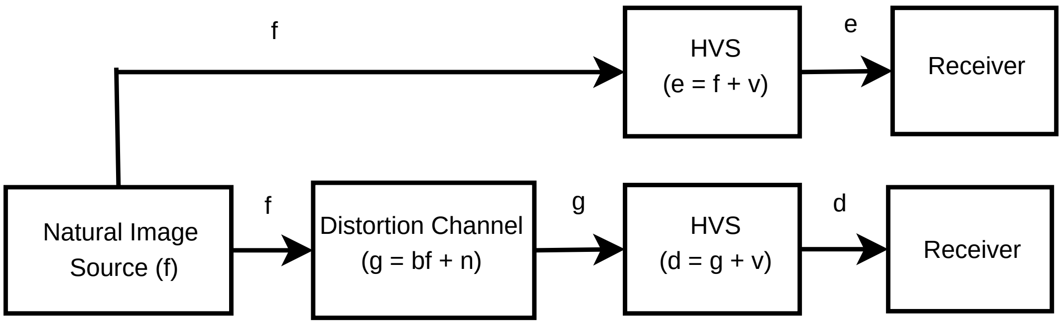


Figure 2.3: Block diagram of VIF quality assessment system

The quality of the image is described using error metrics that quantify the loss of structural information in the image. Illumination does not affect the structure of objects in a scene and the structure comparison is designed to be independent of illumination. Luminance and contrast are computed using the mean and standard deviation of local image patches. The crucial step in the development of SSIM is in defining the structure comparison term, which should capture the structural distortions in an image as seen by the human eye. The correlation or the inner product between mean and variance normalized image patches is used as a simple and effective measure to quantify structural similarity. Despite its simplicity, SSIM correlates extraordinarily well with perceptual image quality. Several improvements to the structural similarity framework have been proposed, including multi-scale structural similarity and translation insensitive Complex Wavelet SSIM (CWSSIM) [18, 19].

A more recent development is the Visual Information Fidelity (VIF) Index. This approach views IQA as an information fidelity problem, as opposed to a signal fidelity problem. The VIF index hypothesizes that visual quality is related to the amount of information that the HVS can extract from an image. Figure 2.3 summarizes the VIF approach. Reference images are assumed to be the output of a natural image source represented using a powerful natural scene statistical (NSS) model [20–23]. The simple, yet powerful NSS model that VIF employs is a Gaussian Scale Mixture (GSM) model [23]. The test image is assumed to be the output of a distortion channel through which the reference image passes. A blur plus additive noise distortion model in the wavelet domain is used as the channel model. Further, the HVS is modeled as a noisy channel, since neural noise and other factors limit the information it can extract from an image. The ratio of the information communicated in the test image channel to that in the reference image channel serves as the quality index. Also, a precursor to VIF, known as the Information Fidelity Criterion (IFC), is described in [24].

2.2 Video Quality Assessment

VQA research has evolved along the same trajectory as IQA. MSE and PSNR are still heavily used due to their simplicity. A large portion of research into video quality metrics over the past twenty years has concentrated on HVS-based quality metrics. A block diagram of a generic HVS-based quality metric is illustrated in Fig. 2.4.

This system is identical to the generic HVS-based IQA system described earlier, except for the block termed “temporal filtering”. It is believed that two kinds of temporal mechanisms exist in the early stages of processing in

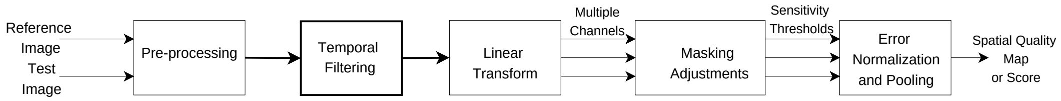


Figure 2.4: Block diagram of HVS-based video quality assessment system

the visual cortex, one lowpass and one bandpass. HVS-based VQA algorithms attempt to model these temporal mechanisms of the HVS in the temporal filtering block. For example, an early HVS-based VQA metric, known as the Moving Pictures Quality Metric (MPQM) [25], analyzed the video signal using a spatial Gabor filterbank and a temporal mechanism consisting of one bandpass and one low-pass filter. Separable spatio-temporal models inadequately describe human visual response; rather, visual motion sensing is better modeled using temporal filters whose response depends on the spatial frequency [25]. Hence, measurement of the CSF as a non-separable function of spatial and temporal frequencies was performed using psychophysical experiments in the develpment MPQM. This metric was improved upon using a more recent model consisting of two IIR filters to model the lowpass and bandpass mechanisms in the HVS [26] to develop the Perceptual Distortion Metric (PDM) [27]. Likewise, the Sarnoff JND vision model was extended to video by including temporal filters similar to those used in PDM [9]. Watson proposed a computationally efficient VQA algorithm known as the Digital Video Quality (DVQ) Metric, using the DCT in the linear transform stage and a single-channel IIR temporal mechanism [28]. A more recent scalable wavelet based video distortion metric [29] uses a single channel filter to model the temporal mechanisms in the HVS and an orthonormal Haar wavelet spatial decomposition.

Very simple and preliminary extensions of both SSIM and VIF have been proposed for VQA [30, 31]. For example, [30] explains a simple frame-by-frame SSIM implementation that proved competitive with the proponents in the VQEG Phase I FR-TV tests. The performance of this index was found to be less stable in areas of large motion, since the frame level SSIM index incorporates no motion information. To make the metric more robust, a simple motion weighting adjustment was proposed where motion vectors are computed using block matching [30]. More recently, the SSIM index was used in conjunction with statistical models of visual speed perception [32]. This method applies the SSIM index frame by frame on the video and uses motion information in designing weights to combine local SSIM measurements into a single quality score for the entire video sequence.

The VIF index was extended to video by applying the same statistical model used for static images on spatio-temporal derivatives of the reference and test videos and by using the same information-theoretic formulation [31]. However, the accuracy of using NSS models developed for static images in the video scenario is questionable. Nevertheless, this approach has also proved to be competitive with the best algorithms in the VQEG tests.

As an indication of their performance, three of the indices mentioned above, namely the Sarnoff JND metric, PDM and DVQ were proponents in the evaluation conducted by the Video Quality Experts Group (VQEG) as part of their Phase 1-FRTV study in 2000 [2]. This study concluded that the performance of all the proponents were statistically equivalent and that the performance of all models were statistically equivalent to PSNR! HVS-based video quality metrics suffer from the same drawbacks that were pointed out earlier for HVS-based image quality metrics. Additionally, they suffer from inaccurate modeling of the temporal mechanisms in the HVS. All the metrics mentioned above use either one or two temporal channels and model the temporal tuning of the neurons in area V1 of the visual cortex only. This is insufficient as it is well known that area MT of the extra-striate cortex plays an important role in motion perception. More recently, the response of neurons in area MT have been studied and models of motion sensing in the human eye have been proposed [33, 34]. To the best of my knowledge, no HVS-based quality metric incorporates these models to account for the second stage of motion processing in area MT of the HVS.

Thus, in recent years, there has been an increased interest in models that describe the distortions in the video sequence that the human eye is sensitive to and that equate with loss of quality; for example, blurring, blocking artifacts, fidelity of edge and texture information in the signal, color information, contrast and luminance of registered patches in the spatial and frequency domain etc. The VQEG conducted another study in 2003, labeled Phase-II FR-TV study, to obtain finer discrimination between models than the Phase-I study [35]. Five of the six proponent models tested by the VQEG in its Phase II testing utilized feature vectors such as those described above in predicting quality [35]. Although the proponent models performed better in this study than in Phase-I, the Phase-II study emphasized a specific, and hence limited application domain, focusing on digitally encoded television. One of the prominent VQA algorithms that belongs to this class and was a proponent in the Phase II study is the Video Quality Metric (VQM) developed at the National Telecommunications and Information Administration (NTIA) [36]. VQM has been standardized by the American National Standards Institute (ANSI) and has been included as a normative method in two International Telecommunications Union (ITU) recommendations.

2.3 Conclusion

There is a need for improvement in the performance of objective quality metrics for video. Most current metrics are benchmarked using the metrics in Phase I of VQEG testing, which have been shown to be statistically equivalent to PSNR in the study. This indicates the potential for improvement in the performance of video quality metrics. Most of the metrics proposed in the literature have been simple extensions of quality metrics for images. Biological vision systems devote considerable resources to motion processing. Presentation of video sequences to human subjects induces visual experience of motion and perceived distortion in video sequences is a combination of both spatial and motion artifacts. For example, motion artifacts such as ghosting, jitter etc. are clearly visible in video signals distorted by compression. Thus, VQA is not a straight forward extension of IQA. I believe that metrics specific to video, and that incorporate modeling of motion as well as temporal distortions in video, need to be developed for accurate quality prediction. To date, there has been little work done in these directions which greatly motivates my work.

Chapter 3

Unified Treatment of Full Reference Image Quality Assessment Algorithms

Perceptual Quality Assessment (QA) algorithms predict the subjective visual quality of an image or video sequence. Full reference QA algorithms assume the availability of a "perfect" quality reference signal and have contributed to a large body of work on QA. Several different full reference image QA algorithms exist in the literature. These algorithms differ in a variety of ways, including the philosophy behind the design of the algorithm, the computational complexity, the presumed viewing conditions incorporated in the model, and the format of the output (local quality measures for different regions of the image, a single global quality measure for the entire image, probability of discrimination between the reference and test images). What, then, is the right algorithm to use in a given application? Typically, this question has been addressed by comparing the performance of each algorithm against human subjective evaluation of visual quality. A database of distorted images is constructed and ground truth quality scores for these images are collected from human observers. To assess the performance of a QA system, algorithm predictions are fitted to the subjective scores using a monotonic function and comparisons are made using statistical tests such as the correlation coefficient, Spearman Rank Order Correlation Coefficient (SROCC), or Root Mean Squared Error (RMSE) between ground truth scores and algorithm predictions.

Such an analysis, while valuable, may present certain drawbacks and need not necessarily correlate well with the performance of a QA system. First of all, the results obtained are only indicative of the performance of the QA system *on the particular database* of images used in the performance evaluation. The performance indices obtained from such a study cannot necessarily be used to predict the performance of the same QA system on other databases containing images having different content, different distortion types and distortion strengths, different experimental setups such as viewing distances or display devices and so on. Additionally, the results depend on the form of the fitting function that is used to map algorithm predictions to human Mean Opinion Scores (MOS) and the optimization algorithm used to determine the parameters of the fitting function. Finally, most QA systems have several free parameters which, when varied, result in different performance indices even while testing on the same database.

I envision that a suitable analysis of different QA indices in a general mathematical framework may deepen our understanding of these indices and enable a unification of ideas derived from different first principles. In this chapter, I study two recently developed, popular image QA paradigms - the Structural SIMilarity paradigm (SSIM) [15, 17] and the Visual Information Fidelity (VIF) paradigm [24, 37] in a general probabilistic framework. I attempt to relate the SSIM and VIF QA paradigms to each other as well as to traditional QA indices: the Mean Squared Error (MSE) and perception based image QA algorithms.

In this dissertation, I use “QA metric” and “QA algorithm” interchangeably, although the output of a QA algorithm is not truly a metric in the mathematical sense of the word. However, due to its rampant use in QA literature, I also use the term “metric” acknowledging my own abuse of terminology. In Section 3.1, I explain the notation that will be used throughout this chapter. In Section 3.2, I describe the SSIM paradigm for image QA and I relate structural similarity metrics to more traditional approaches to QA that have dominated much of the research on this problem for the past two or three decades: the Mean Squared Error (MSE) and Human Visual System (HVS) modeling based methods. In particular, I express the SSIM index as an MSE between certain normalized variables. I also demonstrate that the SSIM index performs a contrast masking normalization, similar to HVS based QA metrics. In Section 3.3, I study the VIF paradigm for image QA. I show that a precursor to the VIF index, known as the Information Fidelity Criterion (IFC), is equivalent to the structure term of the SSIM index applied in the sub-band filtered domain. I also study the relation of the more sophisticated VIF index to the SSIM index.

3.1 Notation

I begin by introducing the notation that I will use throughout this chapter. Let $F(\mathbf{i})$ denote a random variable that models a pixel at spatial location $\mathbf{i}$ in the reference image. Similarly, let $G(\mathbf{i})$ denote a random variable that models the corresponding pixel from the test image. Let $\tilde{f}(\mathbf{i})$ and $\tilde{g}(\mathbf{i})$ denote the reference and test images respectively. Define two sequences of vectors $\mathbf{f}(\mathbf{i})$ and $\mathbf{g}(\mathbf{i})$ of dimension N, where $\mathbf{f}(\mathbf{i})$ is composed of N elements of $\tilde{f}(\mathbf{i})$ spanned by a window B_1 and similarly for $\mathbf{g}(\mathbf{i})$. Thus, if the window B_1 is specified by a set of relative indices, then $\mathbf{f}(\mathbf{i}) = \{\tilde{f}(\mathbf{i}+\mathbf{j}), \mathbf{j} \in B_1\}$. To index each element of $\mathbf{f}(\mathbf{i})$, I use the notation $\mathbf{f}(\mathbf{i}) = [f_1(\mathbf{i}), f_2(\mathbf{i}), \ldots, f_N(\mathbf{i})]^T$. Although the window B_1 can be of any shape, in practice, it usually spans a rectangular region of connected pixels. Consider the linear shift-invariant filtering of $f(\mathbf{i})$ and $g(\mathbf{i})$ by a family of two-dimensional sub-band kernels, denoted $h(\mathbf{i}, k)$, where k indexes over each filter in the family. Let $X(\mathbf{i}, k)$ and $Y(\mathbf{i}, k)$ denote random variables that model the coefficient at spatial location $\mathbf{i}$ obtained by filtering the reference and test image patches with the k^{th} filter $h(\mathbf{i}, k)$ respectively. I will also be interested in random vectors defined by a collection of these coefficients. Define the M-dimensional random vector $\mathbf{X}(\mathbf{i}, k)$ that contains M coefficients of $X(\mathbf{i}, k)$ spanned by a window B_2. A similar definition applies for $\mathbf{Y}(\mathbf{i}, k)$. Let $\tilde{x}(\mathbf{i}, k)$ and $\tilde{y}(\mathbf{i}, k)$ denote the coefficients of the k^{th} sub-band of the reference and test images, respectively. Finally, define M dimensional vectors $\mathbf{x}(\mathbf{i}, k)$ and $\mathbf{y}(\mathbf{i}, k)$ that contain M coefficients of $\tilde{x}(\mathbf{i}, k)$ and $\tilde{y}(\mathbf{i}, k)$ spanned by B_2 respectively, i.e. $\mathbf{x}(\mathbf{i}, k) = \{\tilde{x}(\mathbf{i}+\mathbf{j}, k), \mathbf{j} \in B_2\}$ and similarly for $\mathbf{y}(\mathbf{i}, k)$. I use similar notation to index each element of $\mathbf{x}(\mathbf{i}, k)$ and $\mathbf{y}(\mathbf{i}, k)$, i.e. $\mathbf{x}(\mathbf{i}, k) = [x_1(\mathbf{i}, k), x_2(\mathbf{i}, k), \ldots, x_M(\mathbf{i}, k)]^T$ etc.

3.2 Structural Similarity Metrics

3.2.1 The SSIM Index

A new philosophy for image QA based on the measurement of *structural information* in an image was proposed in [15, 17], which has since received significant visibility in the research community, in addition to widespread adoption in the image and video industry. The SSIM philosophy attempts to avoid the drawbacks of the traditional error sensitivity philosophy that motivated many earlier HVS based QA models [1, 7, 10, 28, 38]. The structural similarity paradigm hypothesizes that the visual quality of a given image is related to the loss of structural information with respect to the reference [15, 17]. The structure of objects in a scene is presumed to be independent of the illumination of the scene. Hence, the effects of illumination (luminance and contrast) are ignored in defining the structural content of a scene. Since a QA index is intended to predict human performance, the SSIM index tacitly assumes that subjective evaluation can be separated into three corresponding tasks - luminance comparison, contrast comparison and structure comparison. All of these comparisons are carried out *locally* over image patches.

The luminance of an image patch $\mathbf{f}(\mathbf{i})$ is estimated as its mean intensity

$$\mu_{\mathbf{f}(\mathbf{i})} = \frac{1}{N} \sum_{j=1}^{N} f_j(\mathbf{i})$$

The contrast of an image patch is estimated as the standard deviation of the patch

$$\sigma^2_{\mathbf{f}(\mathbf{i})} = \frac{1}{N}\sum_{j=1}^{N}\left(f_j(\mathbf{i}) - \mu_{\mathbf{f}(\mathbf{i})}\right)^2 \tag{3.1}$$

The luminance comparison function $l[\mathbf{f}(\mathbf{i}), \mathbf{g}(\mathbf{i})]$ and the contrast comparison function $c[\mathbf{f}(\mathbf{i}), \mathbf{g}(\mathbf{i})]$ between image patches $\mathbf{f}(\mathbf{i})$ from the reference image and $\mathbf{g}(\mathbf{i})$ from the test image are then defined as

$$l[\mathbf{f}(\mathbf{i}), \mathbf{g}(\mathbf{i})] = \frac{2\mu_{\mathbf{f}(\mathbf{i})}\mu_{\mathbf{g}(\mathbf{i})} + C_1}{\mu^2_{\mathbf{f}(\mathbf{i})} + \mu^2_{\mathbf{g}(\mathbf{i})} + C_1} \tag{3.2}$$

$$c[\mathbf{f}(\mathbf{i}), \mathbf{g}(\mathbf{i})] = \frac{2\sigma_{\mathbf{f}(\mathbf{i})}\sigma_{\mathbf{g}(\mathbf{i})} + C_2}{\sigma^2_{\mathbf{f}(\mathbf{i})} + \sigma^2_{\mathbf{g}(\mathbf{i})} + C_2} \tag{3.3}$$

The constants C_1, C_2 are added to prevent numerical instability when the denominators are small. The structure comparison is performed after normalizing the image patches for mean luminance and contrast and the structure comparison function $s[\mathbf{f}(\mathbf{i}), \mathbf{g}(\mathbf{i})]$ is given by

$$s[\mathbf{f}(\mathbf{i}), \mathbf{g}(\mathbf{i})] = \frac{\sigma_{\mathbf{f}(\mathbf{i})\mathbf{g}(\mathbf{i})} + C_3}{\sigma_{\mathbf{f}(\mathbf{i})}\sigma_{\mathbf{g}(\mathbf{i})} + C_3} \tag{3.4}$$

where the sample covariance between $\mathbf{f}(\mathbf{i})$ and $\mathbf{g}(\mathbf{i})$ is

$$\sigma_{\mathbf{f}(\mathbf{i})\mathbf{g}(\mathbf{i})} = \frac{1}{N}\sum_{j=1}^{N}\left(f_j(\mathbf{i}) - \mu_{\mathbf{f}(\mathbf{i})}\right)\left(g_j(\mathbf{i}) - \mu_{\mathbf{g}(\mathbf{i})}\right) \tag{3.5}$$

The sample standard deviation and covariance as defined in (3.1) and (3.5) differ slightly from the original definition in [15]. The estimates in (3.1) and (3.5) correspond to the moment estimates or the Maximum Likelihood

(ML) estimates assuming a Gaussian distribution on the quantities [39], as opposed to the unbiased estimates used in [15]. I have defined these terms here with N in the denominator, as opposed to $N-1$, to avoid inconvenient notation and better clarity of analysis. This modification does not affect estimates of the quantities significantly. Finally, the SSIM index between image patches $\mathbf{f}(\mathbf{i})$ and $\mathbf{g}(\mathbf{i})$ is defined as

$$\text{SSIM}[\mathbf{f}(\mathbf{i}), \mathbf{g}(\mathbf{i})] = l[\mathbf{f}(\mathbf{i}), \mathbf{g}(\mathbf{i})] \,.\, c[\mathbf{f}(\mathbf{i}), \mathbf{g}(\mathbf{i})] \,.\, s[\mathbf{f}(\mathbf{i}), \mathbf{g}(\mathbf{i})] \tag{3.6}$$

The original metric based on structural similarity, known as the Universal Quality Index (UQI), is also defined by (3.6), with $C_1 = C_2 = C_3 = 0$. The constants C_1, C_2, C_3 were included to stabilize the (renamed) SSIM index to avoid instability when the denominators of the luminance, contrast and structure comparison terms became too small. Although the SSIM index is defined by three terms, the structure term in the SSIM index is generally regarded as the most important, since variations in luminance and contrast of an image do not affect visual quality as much as structural distortions [17]. Moreover, most commonly occurring distortions (with the exception of exposure correction) modify the local average luminance only slightly. Since I will be dealing with the structure term of the SSIM index without the constant C_3 repeatedly throughout this chapter, I use the following notation to describe the normalized covariance or the structure term.

$$\hat{\rho}[\mathbf{f}(\mathbf{i}), \mathbf{g}(\mathbf{i})] = \frac{\sigma_{\mathbf{f}(\mathbf{i})\mathbf{g}(\mathbf{i})}}{\sigma_{\mathbf{f}(\mathbf{i})}\sigma_{\mathbf{g}(\mathbf{i})}} \tag{3.7}$$

The reason for this notation will be obvious from the discussion in Section 3.2.2.

3.2.2 Probabilistic SSIM Index

I now describe the SSIM index in a probabilistic framework. First of all, observe that $\hat{\rho}[\mathbf{f}(\mathbf{i}), \mathbf{g}(\mathbf{i})]$ is the square root of the coefficient of determination in a simple linear regression model between $\{f_j(\mathbf{i}), 1 \leq j \leq N\}$ and $\{g_j(\mathbf{i}), 1 \leq j \leq N\}$ [40]. Regression analysis assumes that the independent or regressor variable (the reference pixels in my case) is controllable, while $g_j(\mathbf{i})$ are samples of a random variable that I denote by $G(\mathbf{i})$. However, since I am interested in assessing the quality of any given image amongst infinite possibilities, it would be fair to assume that the regressor is not controllable and that the reference image pixels $f_j(\mathbf{i})$ are samples of a random variable $F(\mathbf{i})$. Such an assumption corresponds to correlation analysis [40] and $\hat{\rho}[\mathbf{f}(\mathbf{i}), \mathbf{g}(\mathbf{i})]$ is the ML estimate of the correlation coefficient between $F(\mathbf{i})$ and $G(\mathbf{i})$ under the assumption that $[F(\mathbf{i}), G(\mathbf{i})]$ are jointly Gaussian. Based on this observation, define a *probabilistic* SSIM index between random variables $F(\mathbf{i})$ and $G(\mathbf{i})$, whose structure term is given by:

$$\rho[F(\mathbf{i}), G(\mathbf{i})] = \frac{\mathrm{Cov}[F(\mathbf{i}), G(\mathbf{i})]}{\sqrt{\mathrm{Var}[F(\mathbf{i})]}\sqrt{\mathrm{Var}[G(\mathbf{i})]}} \tag{3.8}$$

which is the correlation coefficient between the random variables.

The distinction between the structure terms of the probabilistic SSIM index defined by (3.8) and the sample SSIM index defined by (3.7) is quite significant. Notice that (3.7) coincides with the ML estimate of the correlation coefficient only under the assumption that $[F(\mathbf{i}), G(\mathbf{i})]$ are jointly Gaussian. In all other cases, the estimate of the correlation coefficient as defined by (3.7) is inaccurate and does not incorporate any information regarding the distribution of the reference image pixels. Additionally, note that the use of the correlation coefficient in the SSIM index implies measurement of the degree of *linear dependence* between the reference and test images as a measure of visual quality.

3.2.3 Relation to MSE

In this section, I will relate the structure term in the SSIM index to MSE. The MSE between image patches $\mathbf{f}(\mathbf{i})$ and $\mathbf{g}(\mathbf{i})$ is

$$\mathrm{MSE}[\mathbf{f}(\mathbf{i}), \mathbf{g}(\mathbf{i})] = \frac{1}{N}\sum_{j=1}^{N}[f_j(\mathbf{i}) - g_j(\mathbf{i})]^2$$

Define normalized random variables

$$F'(\mathbf{i}) = \frac{F(\mathbf{i}) - \mathbf{E}[F(\mathbf{i})]}{\sqrt{\mathrm{Var}[F(\mathbf{i})]}} \tag{3.9}$$

$$G'(\mathbf{i}) = \frac{G(\mathbf{i}) - \mathbf{E}[G(\mathbf{i})]}{\sqrt{\mathrm{Var}[G(\mathbf{i})]}} \tag{3.10}$$

where $\mathbf{E}$ stands for the expectation operator. Observe that:

$$\mathbf{E}\big\{[F'(\mathbf{i}) - G'(\mathbf{i})]^2\big\} = 2\,\{1 - \rho[F(\mathbf{i}), G(\mathbf{i})]\} \tag{3.11}$$

It is straightforward to show that the relation (3.11) holds for the estimates of the correlation coefficient and MSE as well, assuming once again that $[F(\mathbf{i}), G(\mathbf{i})]$ are jointly Gaussian:

$$\begin{aligned} \mathrm{MSE}\left(\frac{\mathbf{f}(\mathbf{i}) - \mu_{\mathbf{f}(\mathbf{i})}}{\sigma_{\mathbf{f}(\mathbf{i})}}, \frac{\mathbf{g}(\mathbf{i}) - \mu_{\mathbf{g}(\mathbf{i})}}{\sigma_{\mathbf{g}(\mathbf{i})}}\right) &= 2\left(1 - \frac{\sigma_{\mathbf{f}(\mathbf{i})\mathbf{g}(\mathbf{i})}}{\sigma_{\mathbf{f}(\mathbf{i})}\sigma_{\mathbf{g}(\mathbf{i})}}\right) \\ &= 2[1 - \hat{\rho}(\mathbf{f}(\mathbf{i}), \mathbf{g}(\mathbf{i}))] \end{aligned} \tag{3.12}$$

Thus, the structure term in the SSIM index essentially computes an MSE between image patches, after normalizing them for their mean and standard deviations. This is not surprising in view of the fact that the structure term of the SSIM index is defined to be independent of the mean and standard deviation of the image intensity values. However, this observation would prove valuable in *optimizing* image processing algorithms for the SSIM index, since optimization with respect to MSE is a well studied and tractable problem. Indeed, recent work studies optimization of de-noising and other algorithms by minimizing the SSIM index between the de-noised and original images, as opposed to traditional techniques that minimize the MSE between these [41–44].

3.2.4 Relation to HVS Based Metrics

HVS based metrics use psychophysical measurements of the characteristics of the vision system to compute visual quality [38]. QA models based on the HVS are, in general, rather elaborate and model several different aspects of the HVS such as luminance masking, contrast sensitivity and resolution drop-off with eccentricity, as well as aspects of viewing conditions such as viewing distance and display device characteristics. Here, I only consider modeling of the contrast masking property of the HVS. Specifically, I show that the structure term of the SSIM index is equivalent to certain contrast masking models of the HVS. Contrast masking refers to the reduction in visibility of a signal component due to the presence of another signal component of similar frequency and orientation in a local spatial neighborhood. In the context of IQA, the presence of large signal energy in the image content (masker) *masks* the visibility of noise or distortions (target) in these regions.

Contrast masking has been modeled in a variety of ways in the literature. The Daly model uses a threshold elevation approach to model contrast masking [10]. Threshold elevation refers to the difference in contrast at which an observer is able to distinguish between a masker of a certain contrast and a signal plus masker, when both signals are identical except for their contrast. Threshold elevation has been studied extensively [45–47] and models obtained therein are used to normalize the differences between the reference and test signal. The disadvantage of a threshold elevation approach is that it is less suitable in the case of supra-threshold distortions, although it works very well in predicting whether an observer can simply discriminate between the reference and test images.

Contrast masking has also been modeled using contrast gain control models, which generalize better to supra-threshold distortions. Gain control models a mechanism that allows a neuron in the HVS to adjust its response to the ambient contrast of the stimulus, thereby keeping the neural responses within their permissible dynamic range [48]. The contrast response function, which relates the response of a neuron to the input contrast, is modeled using a non-linear function that is expansive at lower contrasts and compressive at high contrasts [12, 49–52]. This model usually takes the form of a divisive normalization, where the response of a neuron has an accelerating nonlinearity, but is also inhibited divisively by the response of a local pool of neurons plus a saturation constant. It has been suggested that divisive normalization results in efficient encoding since it eliminates the statistical dependencies that are present when typical natural images are decomposed using linear filters [52]. The saturation constant determines the range of contrasts that the neuron is responsive to and is important for a number of reasons. First of all, it prevents division by zero. Secondly, there is a range of very low masking contrasts (lower than the threshold of detection of the stimulus) for which the masker has little or no effect in detecting the target. In fact, several studies report that when the contrast is close to the threshold of detection of the target, the presence of a mask may in fact *facilitate* the detection of the target [45, 47, 50]. The saturation constant explains the response in this regime where the masker contrast is close to the baseline contrast sensitivity threshold.

Different models for contrast gain control exist in the literature that share similar properties as outlined above. For brevity, I only discuss some models that have been used in the IQA framework. HVS based metrics typically decompose the image using a linear sub-band decomposition, and masking is modeled in the sub-band decomposed domain. Teo and Heeger [12] use the following gain control model to define the response $R[\tilde{x}(\mathbf{i}, k)]$ of a neuron with input $\tilde{x}(\mathbf{i},k)$:

$$R[\tilde{x}(\mathbf{i},k)] = \kappa \frac{\tilde{x}(\mathbf{i},k)^2}{\sum_{k \in K_1} \tilde{x}(\mathbf{i},k)^2 + C} \tag{3.13}$$

Here, κ and C restrict the dynamic range of the response and C is a saturation constant. Summation over the sub-bands in the denominator is only carried out over those sub-bands with the same frequency, but different orientations. A related model has also been proposed by Watson *et. al.* in [51].

Safranek and Johnson [53] use the following normalization:

$$R[\tilde{x}(\mathbf{i},k)] = \frac{\tilde{x}(\mathbf{i},k)}{\max\{1, \left[\sum_{k \in K_2} \alpha_k \tilde{x}(\mathbf{i},k)\right]^{\gamma}\}}$$

where α_k are weights determined based on HVS measurements and γ is a constant. In this model, summation over the sub-bands in the denominator is carried out over all sub-bands except the DC band. Instead of an additive saturation constant, the Safranek-Johnson model uses the maximum to account for low signal contrast regions.

Lubin [7] uses a sigmoid nonlinearity to model masking:

$$R[\tilde{x}(\mathbf{i},k)] = \frac{|\tilde{x}(\mathbf{i},k)|}{1 + \kappa|\tilde{x}(\mathbf{i},k)|^{\alpha} + |x(\mathbf{i},k)|^{\gamma}}$$

where κ,α,γ are constants.

It is evident that the definition of the normalized variables in (3.9),(3.10) is very similar to divisive normalization models of contrast gain control in HVS based metrics. The SSIM contrast masking model can be defined by:

$$R[f_j(\mathbf{i})] = \frac{f_j(\mathbf{i}) - \mu_{\mathbf{f}(\mathbf{i})}}{\sqrt{\frac{1}{N}\sum_{j=1}^{N}\left[f_j(\mathbf{i}) - \mu_{\mathbf{f}(\mathbf{i})}\right]^2}} \tag{3.14}$$

HVS based QA systems compute a Minkowski error between the outputs of the contrast gain control model (as well as models of other aspects of the HVS incorporated in QA) to the reference and test image patches as an index of quality, often with a Minkowski exponent of 2 [7, 12, 38]. Similarly, observe that the structure term of SSIM in (3.11) is a monotonic function of the square of the Minkowski error between the outputs of the SSIM contrast gain control model in (3.14) with exponent 2.

Relating the SSIM metrics to contrast masking models used in HVS based QA algorithms provides insights on the need for the constant C_3 in the denominator of (3.4). C_3 is added to the numerator of (3.4) only to ensure that the structure term is always bounded to lie between 0 and 1 and is not as important as the constant in the denominator. C_3 in SSIM plays a similar role as the saturation constants in HVS based metrics, since contrast masking effects are minimal in low signal energy regions. This is the reason that the performance of the SSIM index is superior to the UQI index, as demonstrated in [15]. I could define the contrast gain control model with a saturation constant as

$$R[f_j(\mathbf{i})] = \frac{f_j(\mathbf{i}) - \mu_{\mathbf{f}(\mathbf{i})}}{\sqrt{\frac{1}{N}\sum_{j=1}^{N}\left[f_j(\mathbf{i}) - \mu_{\mathbf{f}(\mathbf{i})}\right]^2 + \kappa}} \tag{3.15}$$

Using this gain control model to compute the MSE between normalized variables, as opposed to (3.12), yields

$$\text{MSE}\big\{R[\mathbf{f}(\mathbf{i})], R[\mathbf{g}(\mathbf{i})]\big\} = \frac{\sigma^2_{\mathbf{f}(\mathbf{i})}}{(\sigma_{\mathbf{f}(\mathbf{i})}+\kappa)^2} + \frac{\sigma^2_{\mathbf{g}(\mathbf{i})}}{(\sigma_{\mathbf{g}(\mathbf{i})}+\kappa)^2} - 2\frac{\sigma_{\mathbf{f}(\mathbf{i})\mathbf{g}(\mathbf{i})}}{(\sigma_{\mathbf{f}(\mathbf{i})}+\kappa)(\sigma_{\mathbf{g}(\mathbf{i})}+\kappa)} \tag{3.16}$$

Note that the the constant κ in (3.16) plays the same role as C_3 in (3.4). The choice of adding C_3 to the product of $\sigma_{\mathbf{f}(\mathbf{i})}$ and $\sigma_{\mathbf{g}(\mathbf{i})}$ was made by the designers of the SSIM index to account for the observed effects of instability in low signal energy regions. They may as well have chosen to add a constant κ to each of $\sigma_{\mathbf{f}(\mathbf{i})}$ and $\sigma_{\mathbf{g}(\mathbf{i})}$ separately to account for the same observation, which would have been more consistent with a contrast gain control normalization. The effect of κ is to saturate the contrast masking normalization at low signal energy regions, i.e., when $\sigma_{\mathbf{f}(\mathbf{i})}, \sigma_{\mathbf{g}(\mathbf{i})} \approxeq \kappa$.

The divisive normalization performed by SSIM has several interesting interpretations. The SSIM index uses the standard deviation in a local spatial neighborhood to model inhibition with the mean intensity of the signal fixed at zero, since the mean of the signal is subtracted in the numerator. The use of the standard deviation is related to the definition of the Root Mean Squared (RMS) contrast of an image patch. Indeed, the RMS contrast is defined as the ratio of the sample standard deviation of the intensities to the sample mean. The RMS contrast is generally considered a better measure of contrast for natural images than the bandpass contrast measure that is used in many HVS based models, including the Daly and Lubin models [54, 55]. Additionally, this normalization is consistent with recent studies of the gain control mechanism in the Lateral Geniculate Nucleus (LGN) of cats that show that the gain of a neuron is set by the standard deviation of the intensity values when the mean is fixed [56]. In other words, the measure of contrast that is used by the contrast gain control mechanisms is, in fact, the RMS contrast of the stimulus. Experiments reveal that changes in gain entirely counteract changes in standard deviation, and dividing the standard deviation by a factor multiplies the gain by the same factor [56]. Thus, although the SSIM index was derived from very different first principles, at least part of the reasons for its success can be attributed to similarities it shares with models of the HVS. Of course, I make no claims regarding the suitability of (3.14) as a model for contrast gain control in the HVS. The purpose of this discussion has simply been to draw parallels between the SSIM index and traditional HVS based quality metrics.

3.2.5 Discussion

In this section, I discussed the similarities between the SSIM index and more traditional approaches to QA: MSE and HVS based metrics. I first generalized the concept of structural similarity by defining a probabilistic SSIM index that is "aware" of the underlying statistical distributions of the reference and test pixels. I showed that the structure term of the SSIM index is equivalent to the MSE between the variables after normalizing for their mean and variances. A different relation between SSIM and the MSE between the original variables ($\mathbf{f}(\mathbf{i})$ and $\mathbf{g}(\mathbf{i})$) has been reported [57]. My analysis describes the relation between SSIM and the MSE between *normalized* variables, a distinction that is significant since I attempt to cast the role of this normalization as accounting for contrast masking effects in the HVS.

I also showed that the structure term in the SSIM index is similar to certain models of contrast gain control mechanisms in the HVS. Such an interpretation explains the need for the constants in the definition of the SSIM index. The constants cause saturation of neuronal responses when the contrast of the image is very small and masking effects do not occur. The need for this constant is not explained adequately in the current design philosophy of the SSIM index [15] since the only explanation provided is that it helps avoid numerical instabilities. It has been previously observed that the standard deviations of the reference and test image patches in the denominator of (3.7) reflect masking and that the constant C_3 attempts to account for visibility of distortions when these standard deviations are small [58]. However, the study in [58] was not supported by any analysis. My analysis supports these observations and explicitly links SSIM and contrast masking models.

It is important to note that the SSIM indices perform the gain control normalization in the *image pixel* domain. However, contrast masking in the HVS is a phenomenon that occurs in a frequency-orientation decomposed domain. For example, the masking effect is maximum when the orientation of the masker and the target are parallel and decreases when their orientations are perpendicular [59]. Many divisive normalization models can account for this behavior, since divisive normalization is modeled *after* decomposition using linear filters that are frequency and orientation selective. Thus, the pool of inhibitory neurons include the outputs of linear filters at different orientations and the suppressive weights of neuronal responses in the divisive pool are higher when their orientation is close to the orientation of the neuron whose response is being modeled and lower when the orientation is orthogonal [52, 53]. However, the SSIM metric will not be able to account for such effects. The analysis here suggests that applying the SSIM index in the sub-band filtered domain would result in better performance. Improved versions of the SSIM index that use such frequency decomposition have been proposed [18, 19] and my analysis of the SSIM index within a contrast gain control framework helps us understand the reasons for the improved performance of these metrics. The Complex Wavelet SSIM (CW-SSIM) proposed in [19] operates in a scale-space decomposed space, although CW-SSIM was designed for affine invariance. The Multi-Scale SSIM (MS-SSIM) index also decomposes the image into different scales and calibrates the relative importance of different scales to produce an improved SSIM index [18]. However, the decomposition used in MS-SSIM is limited since it uses a simple low pass filtering (using an average filter) and downsampling procedure. This does not achieve a true frequency and orientation decomposition that is required to explain the masking effects in human vision discussed above. In Section 3.3, I discuss the close relation between the information theoretic metrics and multi-scale structural similarity models.

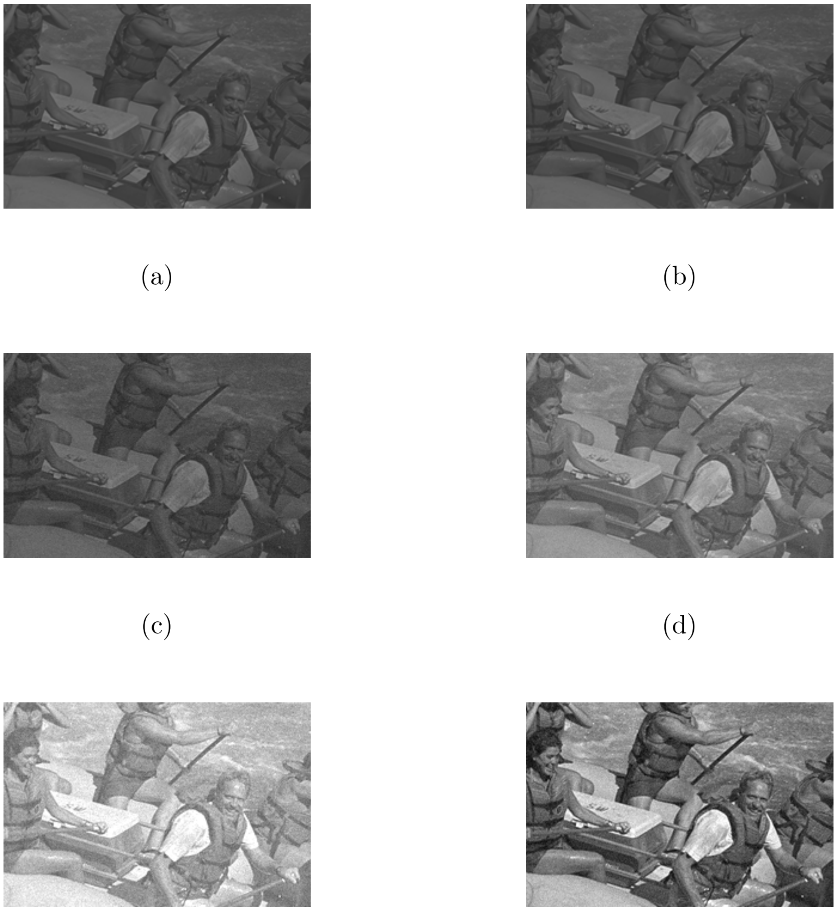
Interestingly, the square of the response of the SSIM contrast gain control model defined by (3.14) is equal to the response of the Teo and Heeger gain control model defined by (3.13) with $\kappa = N$ and $C = 0$, if the same inhibitory pool of neurons is used. This assumes that the coefficients obtained by decomposing the image using a filter bank are zero mean, which is usually true of all sub-bands except the DC band [21, 23].

I briefly discuss the other two terms in the SSIM index - namely the luminance and contrast comparison terms. The SSIM index defines the contrast term using the standard deviation of the pixel values. A more standard definition of contrast, namely the RMS contrast, is defined as the ratio of the standard deviation to the mean of the pixel values. I believe that the contrast comparison term could be better defined using RMS contrast as

$$c[\mathbf{f}(\mathbf{i}), \mathbf{g}(\mathbf{i})] = \frac{2\left(\frac{\sigma_{\mathbf{f}(\mathbf{i})}}{\mu_{\mathbf{f}(\mathbf{i})}}\right)\left(\frac{\sigma_{\mathbf{g}(\mathbf{i})}}{\mu_{\mathbf{g}(\mathbf{i})}}\right) + C_2}{\left(\frac{\sigma_{\mathbf{f}(\mathbf{i})}}{\mu_{\mathbf{f}(\mathbf{i})}}\right)^2 + \left(\frac{\sigma_{\mathbf{g}(\mathbf{i})}}{\mu_{\mathbf{g}(\mathbf{i})}}\right)^2 + C_2} \tag{3.17}$$

It has been found that the luminance and RMS contrast of natural images are statistically independent and that the adaptive gain control mechanisms in the HVS for luminance and contrast operate independently of each other [55]. Comparing the luminance and the standard deviation separately in the SSIM index will result in dependencies between these comparisons that are eliminated by using the RMS contrast instead of the standard deviation. Using (3.2) and (3.17) for luminance and contrast comparison would agree both with the statistics of natural scenes and with HVS processing of natural images. I refer to this modification of the SSIM index as the RC-SSIM (RMS contrast SSIM) index.

Luminance and contrast gain control allows neuronal mechanisms to adjust to the ambient levels of luminance and contrast [55, 60, 61]. Contrast gain control is particularly important in QA since it accounts for masking effects that affect the visibility of distortions in an image. However, in addition to masking, the visibility of these ambient levels of luminance and contrast or the ambient illumination needs to be accounted for. To illustrate this effect, Fig. 3.1 shows a reference image patch and several distorted patches, obtained by multiplying the reference patch by a constant and adding white Gaussian noise. The correlation coefficient and the RMS contrast of Fig. 3.1(c) and 3.1(e) have been adjusted to be identical. However, the mean luminance of Fig. 3.1(c) is lower than that of Fig. 3.1(e). Clearly, the visual quality of the two patches are different. Similarly, the correlation coefficient and the mean luminance of Fig. 3.1(d) and 3.1(f) have been adjusted to be identical. However, the RMS contrast of Fig. 3.1(d) is lower than that of Fig. 3.1(f). Again, the difference in visual quality is obvious. The luminance and contrast comparison terms of the SSIM index explain such changes in illumination, although the structure term lies at the heart of the success of the SSIM index in predicting visual quality. For illustrative purposes, I have used image patches that are far bigger than the 11×11 patches that are used to compute the SSIM index in [15]. Such global changes in illumination result in rather drastic changes in visual quality, while the effect is less pronounced in smaller patches. Global illumination changes occur in contrast enhancement and brightness correction, and are not typical of commonly occurring distortions.



(a) (b) (c) (d) (e) (f)

Figure 3.1: Illustration of the need for the luminance and contrast comparison terms.

3.3 The Information Theoretic Metrics

3.3.1 The IFC and VIF Indices

3.3.1.1 The Information Fidelity Criterion

In the information theoretic approach to QA, the test image is assumed to be the result of the reference image passing through a distortion channel, and its visual quality is hypothesized to be related to the *capacity* of this communication channel [24, 37]. The sub-band filtered coefficients of the reference image are modeled as random variables using natural scene statistic models. The preliminary version of the information theoretic framework, known as the Information Fidelity Criterion (IFC) [24], uses the scalar Gaussian Scale Mixture (GSM) model [62, 63] and each scalar coefficient is modeled as a random variable:

$$X(\mathbf{i}, k) = Z(\mathbf{i}, k)U(\mathbf{i}, k)$$

where $Z(\mathbf{i}, k)$ is a random gain field also known as the mixing density and $U(\mathbf{i}, k)$ is assumed to be an Additive White Gaussian Noise (AWGN) field of unit variance. The distortion channel is modeled as

$$Y(\mathbf{i}, k) = \beta(\mathbf{i}, k)X(\mathbf{i}, k) + V(\mathbf{i}, k)$$

where $\beta(\mathbf{i}, k)$ is the deterministic channel gain and $V(\mathbf{i}, k)$ is AWGN of variance $\sigma_v(\mathbf{i}, k)^2$. $V(\mathbf{i}, k)$ and $X(\mathbf{i}, k)$ are assumed to be independent. A nice property of the GSM model that makes it analytically tractable is that $X(\mathbf{i}, k)$ is normally distributed when conditioned on $Z(\mathbf{i}, k)$. This fact is used in the development of the IFC index which is defined as the capacity of the distortion channel, when conditioned on $Z(\mathbf{i}, k)$. Thus, the IFC index between

$X(\mathbf{i},k)$ and $Y(\mathbf{i},k)$ is defined as the mutual information between these random variables conditioned on $Z(\mathbf{i},k)$ [64]:

$$\begin{aligned}\mathrm{IFC}[X(\mathbf{i},k),Y(\mathbf{i},k)] &= \mathrm{I}[X(\mathbf{i},k),Y(\mathbf{i},k)|Z(\mathbf{i},k)]\\ &= \frac{1}{2}\log_2\left(\frac{\beta(\mathbf{i},k)^2Z(\mathbf{i},k)^2+\sigma_v(\mathbf{i},k)^2}{\sigma_v(\mathbf{i},k)^2}\right)\end{aligned} \tag{3.18}$$

Although the analysis of the IFC index in [24] uses the scalar GSM model, the actual implementation uses the more sophisticated vector GSM model [21, 23] to model the distribution of vectors of coefficients obtained from a sub-band decomposition of an image. The vector model improves over the scalar model since it does not make the poor assumption that wavelet coefficients at adjacent locations, scales and orientations are uncorrelated and models the correlations between these coefficients. The generic GSM model defines a probabilistic model for a vector of coefficients that contains a given coefficient and a collection of its neighbors at adjacent spatial locations, orientations and scales [23]. However, the IFC index uses a simpler model where each sub-band is treated independently and a vector of coefficients is defined as a given coefficient in a sub-band and a collection of its spatially adjacent neighbors in the *same sub-band*. Such a collection is identical to the vector $\mathbf{X}(\mathbf{i},k)$ defined in Section 3.1 obtained by applying a window on each sub-band. The GSM model for such a vector is defined by:

$$\mathbf{X}(\mathbf{i},k) = Z(\mathbf{i},k)\mathbf{U}(\mathbf{i},k) \tag{3.19}$$

Here, $Z(\mathbf{i},k)$ is a scalar multiplier and $\mathbf{U}(\mathbf{i},k)$ is a zero mean Gaussian random vector with covariance matrix $\mathbf{C_U}(k)$. Since $\mathbf{C_U}(k)$ is a positive definite covariance matrix, it has an eigen decomposition given by $\mathbf{C_U}(k) = \mathbf{Q}(k)\mathbf{\Lambda}(k)\mathbf{Q}(k)^T$. $\mathbf{\Lambda}(k)$ is a diagonal matrix containing the eigen values $\{\lambda_j(k), 1 \leq j \leq M\}$ of $\mathbf{C_U}(k)$ and $\mathbf{Q}(k)$ is an orthogonal matrix. The distortion channel is given by

$$\mathbf{Y}(\mathbf{i}, k) = \beta(\mathbf{i}, k)\mathbf{X}(\mathbf{i}, k) + \mathbf{V}(\mathbf{i}, k) \tag{3.20}$$

Again, $\beta(\mathbf{i}, k)$ is a deterministic gain and $\mathbf{V}(\mathbf{i}, k)$ is a zero mean AWGN vector with covariance matrix $\sigma_v(\mathbf{i}, k)^2\mathbf{I}$. $\mathbf{V}(\mathbf{i}, k)$ and $\mathbf{X}(\mathbf{i}, k)$ are assumed to be independent. Then, the vector IFC index, where the subscript v is used to denote the vector metric, is given by

$$\begin{aligned}\mathrm{IFC}_v[\mathbf{X}(\mathbf{i}, k), \mathbf{Y}(\mathbf{i}, k)] &= \mathrm{I}[\mathbf{X}(\mathbf{i}, k), \mathbf{Y}(\mathbf{i}, k)|Z(\mathbf{i}, k)] && (3.21)\\ &= \frac{1}{2}\log_2\left(\frac{|\beta(\mathbf{i}, k)^2 Z(\mathbf{i}, k)^2\mathbf{C_U}(k) + \sigma_v(\mathbf{i}, k)^2\mathbf{I}|}{|\sigma_v(\mathbf{i}, k)^2\mathbf{I}|}\right) && (3.22)\end{aligned}$$

$|\mathbf{I}|$ denotes the determinant of a matrix $\mathbf{I}$.

3.3.1.2 The Visual Information Fidelity Criterion

Although the VIF index uses the vector GSM model, I first describe the scalar version of the VIF metric since understanding this metric will prove useful in analyzing the vector model. The VIF model also uses a more sophisticated distortion model. In addition to the gain and additive noise distortion channel, the HVS is itself modeled as a distortion channel that both the reference and distorted images pass through.

$$Y(\mathbf{i}, k) = \beta(\mathbf{i}, k)X(\mathbf{i}, k) + V(\mathbf{i}, k) + W(\mathbf{i}, k) \tag{3.23}$$

$$T(\mathbf{i}, k) = X(\mathbf{i}, k) + W(\mathbf{i}, k) \tag{3.24}$$

Here, $W(\mathbf{i},k)$ is an AWGN field that has a constant variance κ for all sub-bands and models the "neural noise" in HVS. $W(\mathbf{i},k)$ is assumed to be independent of $X(\mathbf{i},k)$ and $N(\mathbf{i},k)$. $T(\mathbf{i},k)$ is the output of the HVS channel that the reference image passes through. The visual quality of the test image is then defined as the ratio of the capacity of the test channel to that of the reference image channel. Hence, the VIF index $\mathrm{VIF}[X(\mathbf{i},k),Y(\mathbf{i},k)]$ is given by

$$\begin{aligned}
\mathrm{VIF}[X(\mathbf{i},k),Y(\mathbf{i},k)] &= \frac{\mathrm{I}[X(\mathbf{i},k),Y(\mathbf{i},k)|Z(\mathbf{i},k)]}{\mathrm{I}[X(\mathbf{i},k),T(\mathbf{i},k)|Z(\mathbf{i},k)]} \\
\mathrm{I}[X(\mathbf{i},k),Y(\mathbf{i},k)|Z(\mathbf{i},k)] &= \frac{1}{2}\log_2\left(\frac{\beta(\mathbf{i},k)^2 Z(\mathbf{i},k)^2+\sigma_v(\mathbf{i},k)^2+\kappa}{\sigma_v(\mathbf{i},k)^2+\kappa}\right) \qquad (3.25) \\
\mathrm{I}[X(\mathbf{i},k),T(\mathbf{i},k)|Z(\mathbf{i},k)] &= \frac{1}{2}\log_2\left(\frac{Z(\mathbf{i},k)^2+\kappa}{\kappa}\right) \qquad (3.26)
\end{aligned}$$

I will now describe the vector VIF model. Modeling of the reference image coefficients is identical to the vector GSM model.

$$\mathbf{X}(\mathbf{i},k) = Z(\mathbf{i},k)\mathbf{U}(\mathbf{i},k) \qquad (3.27)$$

The distortion channels are given by

$$\mathbf{Y}(\mathbf{i},k) = \beta(\mathbf{i},k)\mathbf{X}(\mathbf{i},k) + \mathbf{V}(\mathbf{i},k) + \mathbf{W}(\mathbf{i},k) \qquad (3.28)$$

$$\mathbf{T}(\mathbf{i},k) = \mathbf{X}(\mathbf{i},k) + \mathbf{W}(\mathbf{i},k) \qquad (3.29)$$

Here, $\mathbf{W}(\mathbf{i},k)$ is a zero mean AWGN vector that models the HVS with covariance matrix $\kappa\mathbf{I}$. $\mathbf{W}(\mathbf{i},k)$ is assumed to be independent of $\mathbf{X}(\mathbf{i},k)$ and $\mathbf{N}(\mathbf{i},k)$. $\mathbf{T}(\mathbf{i},k)$ is the output of the HVS channel that the reference image passes through. The vector VIF index is subsequently given by

$$\mathrm{VIF}_v[\mathbf{X}(\mathbf{i},k),\mathbf{Y}(\mathbf{i},k)] = \frac{\mathrm{I}[\mathbf{X}(\mathbf{i},k),\mathbf{Y}(\mathbf{i},k)|Z(\mathbf{i},k)]}{\mathrm{I}[\mathbf{X}(\mathbf{i},k),\mathbf{T}(\mathbf{i},k)|Z(\mathbf{i},k)]}$$

$$\mathrm{I}[\mathbf{X}(\mathbf{i},k),\mathbf{Y}(\mathbf{i},k)|Z(\mathbf{i},k)] = \frac{1}{2}\log_2\left(\frac{|\beta(\mathbf{i},k)^2 Z(\mathbf{i},k)^2\mathbf{C}_\mathbf{U}(k) + (\sigma_v(\mathbf{i},k)^2+\kappa)\mathbf{I}|}{|(\sigma_v(\mathbf{i},k)^2+\kappa)\mathbf{I}|}\right) \tag{3.30}$$

$$= \frac{1}{2}\sum_{j=1}^{M}\log_2\left(1 + \frac{\beta(\mathbf{i},k)^2 Z(\mathbf{i},k)^2\lambda_j(k)}{\sigma_v(\mathbf{i},k)^2+\kappa}\right) \tag{3.31}$$

$$\mathrm{I}[\mathbf{X}(\mathbf{i},k),\mathbf{T}(\mathbf{i},k)|Z(\mathbf{i},k)] = \frac{1}{2}\log_2\left(\frac{|Z(\mathbf{i},k)^2\mathbf{C}_\mathbf{U}(k)+\kappa\mathbf{I}|}{|\kappa\mathbf{I}|}\right) \tag{3.32}$$

$$= \frac{1}{2}\sum_{j=1}^{M}\log_2\left(1 + \frac{Z(\mathbf{i},k)^2\lambda_j(k)}{\kappa}\right) \tag{3.33}$$

3.3.2 Relation of IFC to SSIM

3.3.2.1 Scalar Model

In this section, I explore the relation between the IFC metric and SSIM. First of all, the GSM model used in the information theoretic metrics results in the sub-band coefficients being Gaussian distributed, when conditioned on the mixing density. The linear distortion channel model results in the reference and test image being jointly Gaussian. Recall that this was the assumption made in defining the structure term of the SSIM index using the correlation coefficient in Section 3.2.1. These observations hint at the possibility that the IFC index may be closely related to SSIM. Having established a monotonic relationship between these metrics, I discovered that it is a well known result in the field of statistical inference and information theory. When two variables are jointly Gaussian, the mutual information between them is a function of just the correlation coefficient [65, 66]. Due to the source and channel model assumptions described in Section 3.3.1.2, $[\mathbf{X}(\mathbf{i},k), \mathbf{Y}(\mathbf{i},k)]$ are jointly Gaussian when conditioned on $Z(\mathbf{i},k)$. Hence, the following relation holds [65].

$$\mathrm{I}[\mathbf{X}(\mathbf{i},k), \mathbf{Y}(\mathbf{i},k)|Z(\mathbf{i},k)] = \frac{1}{2}\log_2\left(\frac{1}{1-\rho[X(\mathbf{i},k), Y(\mathbf{i},k)|Z(\mathbf{i},k)]^2}\right) \quad (3.34)$$

I will now show that the relation (3.34) holds for the estimates of these quantities as well; in other words, that the IFC index in [24] and the structure term of the SSIM index in [15] satisfy the same relation. The computation of the IFC index as described by (3.18) depends on the way the parameters $Z(\mathbf{i},k)$ and $\beta(\mathbf{i},k)$ are estimated. The sample IFC index is defined using estimates of these parameters in (3.18). To obtain estimates of $Z(\mathbf{i},k)$ and $\beta(\mathbf{i},k)$, a local neighborhood of coefficients surrounding the spatial location $\mathbf{i}$ is considered [24]. In the IFC framework, each sub-band is treated independently and hence, a local neighborhood is extracted by considering coefficients in the same sub-band at adjacent spatial locations [24]. Consistent with my earlier notation, I denote this local neighborhood extracted using a window B_2 as $\mathbf{x}(\mathbf{i},k) = [x_1(\mathbf{i},k), x_2(\mathbf{i},k), \ldots, x_M(\mathbf{i},k)]^T$ for the reference image coefficients and $\mathbf{y}(\mathbf{i},k) = [y_1((\mathbf{i},k), y_2(\mathbf{i},k), \ldots, y_M(\mathbf{i},k)]^T$ for the test image. Let $\hat{Z}(\mathbf{i},k)$ denote an estimate of $Z(\mathbf{i},k)$ and similarly for $\hat{\beta}(\mathbf{i},k)$. Now, if I assume that both parameters are estimated using the same window B_2, the ML estimate of $\hat{Z}(\mathbf{i},k)$ is given by [62, 63]

$$\hat{Z}(\mathbf{i},k)^2 = \frac{1}{M}\sum_{j=1}^{M} x_j(\mathbf{i},k)^2 = \sigma^2_{\mathbf{x}(\mathbf{i},k)} \quad (3.35)$$

Here, (3.35) follows from the fact that the sub-band coefficients are assumed to be zero-mean. The least squares estimate of $\hat{\beta}(\mathbf{i}, k)$ and the noise variance are obtained using linear regression [24].

$$\hat{\beta}(\mathbf{i}, k) = \frac{\sigma_{\mathbf{x}(\mathbf{i},k)\mathbf{y}(\mathbf{i},k)}}{\sigma^2_{\mathbf{x}(\mathbf{i},k)}} \tag{3.36}$$

$$\hat{\sigma}_v(\mathbf{i}, k)^2 = \sigma^2_{\mathbf{y}(\mathbf{i},k)} - \hat{\beta}(\mathbf{i}, k)\sigma_{\mathbf{x}(\mathbf{i},k)\mathbf{y}(\mathbf{i},k)} \tag{3.37}$$

$$= \sigma^2_{\mathbf{y}(\mathbf{i},k)} - \frac{\sigma^2_{\mathbf{x}(\mathbf{i},k)\mathbf{y}(\mathbf{i},k)}}{\sigma^2_{\mathbf{x}(\mathbf{i},k)}} \tag{3.38}$$

Finally, substituting the estimates from (3.35) and (3.36) into (3.18), it follows that the sample IFC index is given by

$$\mathrm{IFC}[\mathbf{x}(\mathbf{i}, k), \mathbf{y}(\mathbf{i}, k)] = \frac{1}{2}\log_2\left(\frac{1}{1 - \left[\frac{\sigma_{\mathbf{x}(\mathbf{i},k)\mathbf{y}(\mathbf{i},k)}}{\sigma_{\mathbf{x}(\mathbf{i},k)}\sigma_{\mathbf{y}(\mathbf{i},k)}}\right]^2}\right)$$

$$= \frac{1}{2}\log_2\left(\frac{1}{1 - \hat{\rho}[\mathbf{x}(\mathbf{i}, k), \mathbf{y}(\mathbf{i}, k)]^2}\right) \tag{3.39}$$

Thus, the IFC index at a location in a sub-band is a *monotonic function* of the square of the structure term of the SSIM index (defined by (3.7)) computed at the same location in the same sub-band, as long as the window used in both metrics for estimation purposes are identical. The IFC index is applied in the sub-band filtered domain and is better able to account for the contrast masking properties of the HVS than SSIM (see discussion in Section 3.2.5) and is very closely related to the multi-scale SSIM index in [18]. In fact, my analysis shows that if the same frequency decomposition and estimation windows are used in IFC and multi-scale SSIM, the local quality indices obtained using both metrics would be equivalent due to the monotonic relationship described by (3.39). However, note that no saturation constant appears in (3.39) and therefore, the IFC index will suffer from instability issues in regions where the signal energy in a sub-band is very low, similar to the UQI index. The discussion in Section 3.3.3 will show that the VIF index attempts to compensate for this deficiency.

In the original implementation of the IFC, the estimation windows used for $\hat{Z}(\mathbf{i}, k)$ and $\hat{\beta}(\mathbf{i}, k)$ are both square [24]. However, the size of the window used in estimating the regression coefficients is bigger than that used to estimate the GSM parameter [24]. My analysis assumes that the same window is used in estimating both these parameters. Since the choice of window used in [24] is arbitrary, this is a minor modification of the original framework which does not significantly alter the resulting sample IFC index.

3.3.2.2 Vector Model

I now consider the vector IFC model and explore its relation to the SSIM index. The vector GSM models the joint distribution of vectors of coefficients from the reference image. Thus, the SSIM index needs to be generalized to the case of vector valued random variables by generalizing the definition of the correlation coefficient. In the theory of multi-variate statistical analysis, this is accomplished using canonical correlation analysis [67, 68]. Canonical correlation analysis attempts to find linear combinations of variables in each vector that have maximum correlation. Then, a second set of linear combinations is sought such that the correlation between these is the maximum between all linear combinations that are uncorrelated with the first linear combination, and so on. The canonical correlation coefficients are invariant to linear transformations of the sets of variables. In this section, I explore a natural extension of the notion of the probabilistic SSIM index defined in Section 3.2.2 to vector valued random variables using canonical correlation analysis. I then show that the vector IFC index and the canonical correlation coefficients satisfy a monotonic relationship, thus establishing the equivalence between vector models of the structural similarity and information theoretic paradigms of image QA.

Under the distortion model specified by (3.20) that causes $[\mathbf{X}(\mathbf{i},k),\mathbf{Y}(\mathbf{i},k)]$ to be jointly Gaussian when conditioned on $Z(\mathbf{i},k)$, the mutual information between these random vectors is related to the canonical correlation coefficients [65]:

$$\mathrm{I}(\mathbf{X}(\mathbf{i},k),\mathbf{Y}(\mathbf{i},k)|Z(\mathbf{i},k)) = \frac{1}{2}\sum_{j=1}^{M}\log_2\left(\frac{1}{1-\rho_j[\mathbf{X}(\mathbf{i},k),\mathbf{Y}(\mathbf{i},k)|Z(\mathbf{i},k)]^2}\right) \tag{3.40}$$

where $\rho_j[\mathbf{X}(\mathbf{i},k),\mathbf{Y}(\mathbf{i},k)|Z(\mathbf{i},k)]$ are the canonical correlation coefficients between these variables.

Once again, it is straightforward to show that the relation (3.40) holds for the estimates of these quantities as well. The sample IFC index is computed using estimates $\hat{Z}(\mathbf{i},k)$, $\hat{\mathbf{C}}_{\mathbf{U}}(k)$, $\hat{\beta}(\mathbf{i},k)$ and $\hat{\sigma}_v(\mathbf{i},k)$ of the corresponding quantities in (3.21).

The ML estimates of $\hat{Z}(\mathbf{i},k)$ and $\hat{\mathbf{C}}_{\mathbf{U}}(k)$ used in the IFC index are given by [21, 23, 37]:

$$\hat{\mathbf{C}}_{\mathbf{U}}(k) = \frac{1}{L(k)} \sum_{\mathbf{i} \in A(k)} \mathbf{x}(\mathbf{i}, k)\mathbf{x}(\mathbf{i}, k)^T \tag{3.41}$$

$$\hat{Z}(\mathbf{i}, k)^2 = \frac{\mathbf{x}(\mathbf{i}, k)^T \hat{\mathbf{C}}_{\mathbf{U}}(k)^{-1} \mathbf{x}(\mathbf{i}, k)}{M} \tag{3.42}$$

where $A(k)$ is a set containing the spatial location indices of all vectors in sub-band k, and $L(k)$ is the cardinality of $A(k)$. Let $\hat{\mathbf{C}}_{\mathbf{U}}(k)$ have an eigen decomposition given by $\hat{\mathbf{C}}_{\mathbf{U}}(k) = \hat{\mathbf{Q}}(k)\hat{\mathbf{\Lambda}}(k)\hat{\mathbf{Q}}(k)^T$. $\hat{\mathbf{\Lambda}}(k)$ is a diagonal matrix containing the eigen values $\{\hat{\lambda}_j(k), j = 1, 2, \ldots M\}$ of $\hat{\mathbf{C}}_{\mathbf{U}}(k)$ along the diagonal.

Since $[\mathbf{X}(\mathbf{i}, k), \mathbf{Y}(\mathbf{i}, k)]$ are jointly Gaussian when conditioned on $Z(\mathbf{i}, k)$, the estimates of $\hat{\beta}(\mathbf{i}, k)$ and $\hat{\sigma}_v(\mathbf{i}, k)^2$ obtained using linear regression in [37] coincide with the ML estimates of these quantities. The estimates are given by the expressions in (3.36) and (3.37). Then, the sample IFC index is given by:

$$\mathrm{IFC}_v[\mathbf{x}(\mathbf{i}, k), \mathbf{y}(\mathbf{i}, k)] = \frac{1}{2} \sum_{j=1}^{M} \log_2 \left(1 + \frac{\hat{\beta}(\mathbf{i}, k)^2 \hat{Z}(\mathbf{i}, k)^2 \hat{\lambda}_j(k)}{\hat{\sigma}_v(\mathbf{i}, k)^2}\right) \tag{3.43}$$

Now, I consider estimation of the canonical correlation coefficients. This requires estimates of the covariance matrices of $\mathbf{X}(\mathbf{i}, k)$ and $\mathbf{Y}(\mathbf{i}, k)$ conditioned on $Z(\mathbf{i}, k)$ which I denote as $\hat{\mathbf{C}}_{\mathbf{XX}}(\mathbf{i}, k)$ and $\hat{\mathbf{C}}_{\mathbf{YY}}(\mathbf{i}, k)$, as well as the cross covariance matrix between $\mathbf{X}(\mathbf{i}, k)$ and $\mathbf{Y}(\mathbf{i}, k)$ conditioned on $Z(\mathbf{i}, k)$ denoted as $\hat{\mathbf{C}}_{\mathbf{XY}}(\mathbf{i}, k)$ [68]. Under the GSM model, the reference image coefficients are distributed as zero mean Gaussian random vectors with covariance matrix $Z(\mathbf{i},k)^2\mathbf{C_U}(k)$ when conditioned on the mixing field $Z(\mathbf{i},k)$. Replacing $Z(\mathbf{i},k)^2$ and $\mathbf{C_U}(k)$ with their ML estimates gives

$$\hat{\mathbf{C}}_{\mathbf{XX}}(\mathbf{i},k) = \hat{Z}(\mathbf{i},k)^2\hat{\mathbf{C}}_{\mathbf{U}}(k) \tag{3.44}$$

where the quantities on the RHS are defined by (3.41) and (3.42). Under the assumption that the test image coefficients $\mathbf{Y}(\mathbf{i},k)$ are described by the linear distortion model in (3.28), $\mathbf{Y}(\mathbf{i},k)$ is distributed as a zero mean Gaussian random vector with covariance matrix $\beta(\mathbf{i},k)^2Z(\mathbf{i},k)^2\mathbf{C_U}(k)+\sigma_v(\mathbf{i},k)^2\mathbf{I}$ when conditioned on $Z(\mathbf{i},k)$. The cross covariance matrix between $\mathbf{X}(\mathbf{i},k)$ and $\mathbf{Y}(\mathbf{i},k)$ is $\beta(\mathbf{i},k)Z(\mathbf{i},k)^2\mathbf{C_U}(k)$. Replacing $\beta(\mathbf{i},k)$,$\sigma_v(\mathbf{i},k)$,$Z(\mathbf{i},k)^2$ and $\mathbf{C_U}(k)$ with their ML estimates yields

$$\hat{\mathbf{C}}_{\mathbf{YY}}(\mathbf{i},k) = \hat{\beta}(\mathbf{i},k)^2\hat{Z}(\mathbf{i},k)^2\hat{\mathbf{C}}_{\mathbf{U}}(k) + \hat{\sigma}_v(\mathbf{i},k)^2\mathbf{I} \tag{3.45}$$

$$\hat{\mathbf{C}}_{\mathbf{XY}}(\mathbf{i},k) = \hat{\beta}(\mathbf{i},k)\hat{Z}(\mathbf{i},k)^2\hat{\mathbf{C}}_{\mathbf{U}}(k) \tag{3.46}$$

where (3.36) and (3.37) give estimates of the other quantities in the RHS.

The squares of the canonical correlation coefficients $\hat{\rho}_j(\mathbf{i},k)^2$ are then the eigen values of the matrix $\mathbf{\Sigma}$ given by [68]:

$$\mathbf{\Sigma} = \hat{\mathbf{C}}_{\mathbf{XY}}(\mathbf{i},k)\hat{\mathbf{C}}_{\mathbf{YY}}(\mathbf{i},k)^{-1}\hat{\mathbf{C}}_{\mathbf{XY}}(\mathbf{i},k)\hat{\mathbf{C}}_{\mathbf{XX}}(\mathbf{i},k)^{-1} \tag{3.47}$$

Using the eigen value decomposition of $\hat{\mathbf{C}}_{\mathbf{U}}(k)$ given by $\hat{\mathbf{Q}}(k)\hat{\mathbf{\Lambda}}(k)\hat{\mathbf{Q}}(k)^T$, the canonical correlation coefficients are found to be

$$\hat{\rho}_j(\mathbf{i},k)^2 = \frac{\hat{\beta}(\mathbf{i},k)^2\hat{Z}(\mathbf{i},k)^2\hat{\lambda}_j(k)}{\hat{\beta}(\mathbf{i},k)^2\hat{Z}(\mathbf{i},k)^2\hat{\lambda}_j(k)+\hat{\sigma}_v(\mathbf{i},k)^2}, \quad j = 1,2,\ldots,M \tag{3.48}$$

This yields the monotonic relationship between the vector IFC index and estimates of the squares of the canonical correlation coefficients:

$$\mathrm{IFC}_v[\mathbf{x}(\mathbf{i},k),\mathbf{y}(\mathbf{i},k)] = \frac{1}{2}\sum_{j=1}^{M}\log_2\left(\frac{1}{1-\hat{\rho}_j(\mathbf{i},k)^2}\right) \tag{3.49}$$

Observe that I computed estimates of the canonical correlation coefficients between the reference and test image coefficient vectors, assuming that the reference image coefficients can be described using the GSM model in (3.19) and that the coefficients of the distorted image are described by (3.20). Hence, I have proved that the vector IFC index is equivalent to the (newly defined) *probabilistic* vector SSIM philosophy if the same models are used to describe the distributions of the reference and test image coefficients in both QA systems.

3.3.3 Relation of VIF to SSIM

In this section, I explore the relation of the VIF index to SSIM. From the definition of the VIF index in Section 3.3.1.2, it is apparent that the chief distinction between the IFC and VIF indices is the normalization by the information in the reference image channel in VIF. The distortion channel that the test image passes through in VIF is very similar to the channel in IFC, and the only difference is the addition of a noise component that models human visual processing. I make use of these observations in my analysis of the VIF index below.

3.3.3.1 Scalar Model

The method of estimating $\hat{Z}(\mathbf{i},k)$ in the scalar VIF model is identical to the IFC and is described by (3.35). The VIF distortion model is given by (3.23), which consists of a deterministic gain $\beta(\mathbf{i},k)$ and AWGN $V(\mathbf{i},k)$. The estimates of the gain $\hat{\beta}(\mathbf{i},k)$ and the noise variance $\hat{\sigma}_v(\mathbf{i},k)^2$ are also unchanged from the IFC index and are given by (3.36) and (3.37). The one additional parameter in this model, namely the variance of the HVS noise κ, was hand optimized in [37]. This value is chosen to be 0.1 for all sub-bands, i.e. $\kappa = 0.1$ for all k.

Substituting from (3.35),(3.36) and (3.37) into (3.25) and (3.26), the sample VIF index can be rewritten as

$$\mathrm{VIF}[x(\mathbf{i},k),y(\mathbf{i},k)] = \frac{\log\left[1-\hat{\rho}[\mathbf{x}(\mathbf{i},k),\mathbf{y}(\mathbf{i},k)]^2\left(\frac{\sigma^2_{\mathbf{y}(\mathbf{i},k)}}{\sigma^2_{\mathbf{y}(\mathbf{i},k)}+\kappa}\right)\right]}{\log\left[1-\left(\frac{\sigma^2_{\mathbf{x}(\mathbf{i},k)}}{\sigma^2_{\mathbf{x}(\mathbf{i},k)}+\kappa}\right)\right]} \tag{3.50}$$

Notice that when $\sigma^2_{\mathbf{x}(\mathbf{i},k)}$ is very small, the constant κ dominates the expression in (3.50). I plot the VIF index as a function of $\sigma^2_{\mathbf{x}(\mathbf{i},k)}$ in Fig. 3.2, with $\hat{\rho}[\mathbf{x}(\mathbf{i},k),\mathbf{y}(\mathbf{i},k)] = 0.5$, $\sigma^2_{\mathbf{y}(\mathbf{i},k)} = \sigma^2_{\mathbf{x}(\mathbf{i},k)}$ and $\kappa = 0.1$ as in [37]. The VIF index has large values for very small values of $\sigma^2_{\mathbf{x}(\mathbf{i},k)}$ and this accounts for masking effects in low signal energy regions. Thus, the VIF index attempts to compensate for the lack of a saturation constant in the IFC, which was a drawback of the IFC that was observed in Section 3.3.2.1. The constant κ determines the knee of this curve, and the variance of the reference image coefficients below which the VIF index attempts to compensate for masking effects as discussed in Section 3.2.4.

The VIF index includes a structure comparison term $\hat{\rho}[\mathbf{x}(\mathbf{i},k),\mathbf{y}(\mathbf{i},k)])$, and a contrast comparison term (due to the appearance of functions of $\sigma^2_{\mathbf{y}(\mathbf{i},k)}$ and $\sigma^2_{\mathbf{x}(\mathbf{i},k)}$ in the numerator and denominator respectively), similar to the SSIM index. One of the properties of the VIF index observed in [37] was the fact that it can predict improvement in quality due to contrast enhancement. My analysis of the VIF index explains this effect since the correlation coefficient between a contrast enhanced image and the reference image is 1. The VIF index is > 1 in this case since $\sigma_{\mathbf{y}(\mathbf{i},k)} > \sigma_{\mathbf{x}(\mathbf{i},k)}$ and the contrast comparison in VIF is not symmetric, unlike the contrast comparison in SSIM. Finally, the VIF index avoids certain numerical instabilities that occur in the IFC, since the IFC goes to ∞ as $\hat{\rho}[\mathbf{x}(\mathbf{i},k),\mathbf{y}(\mathbf{i},k)]$ goes to 1. The use of the constant κ in (3.50) ensures that the VIF index is 1 when the reference and test image are identical.

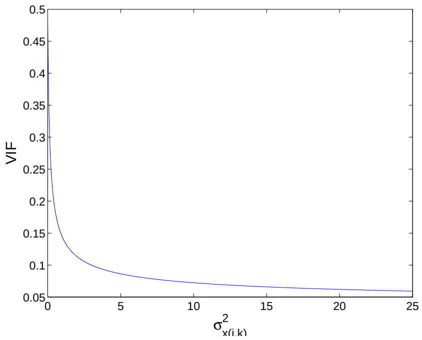


Figure 3.2: Plot of the VIF index as a function of $\sigma^2_{\mathbf{x}(\mathbf{i},k)}$

To better understand the role of these constants, i.e. C_3 in SSIM and

κ in VIF, I perform a sensitivity analysis of SSIM and VIF with respect to the constants. I denote the sensitivity of the structure term of the SSIM index with respect to C_3 as $P_{\text{SSIM}}(C_3)$ and characterize it using partial derivatives as

$$P_{\text{SSIM}}(C_3) = \frac{\partial s[\mathbf{f}(\mathbf{i}), \mathbf{g}(\mathbf{i})]}{\partial C_3} \tag{3.51}$$

$$= \frac{\sigma_{\mathbf{f}(\mathbf{i})}\sigma_{\mathbf{g}(\mathbf{i})} - \sigma_{\mathbf{f}(\mathbf{i})\mathbf{g}(\mathbf{i})}}{(\sigma_{\mathbf{f}(\mathbf{i})}\sigma_{\mathbf{g}(\mathbf{i})} + C_3)^2} \tag{3.52}$$

Firstly, I observe that the sensitivity is maximum when $C_3 = 0$, when the SSIM index reduces to the UQI index. The sensitivity is plotted as a function of C_3 and $\sigma_{\mathbf{f}(\mathbf{i})}\sigma_{\mathbf{g}(\mathbf{i})}$, when the correlation coefficient $\hat{\rho}[\mathbf{f}(\mathbf{i}), \mathbf{g}(\mathbf{i})]$ equals 0.5 in Fig. 3.3. The sensitivity of the SSIM index to the constant is high when both C_3 and $\sigma_{\mathbf{f}(\mathbf{i})}\sigma_{\mathbf{g}(\mathbf{i})}$ are close to 0. This is not surprising in view of the fact that the constant was added in the SSIM index to avoid numerical instabilities when $\sigma_{\mathbf{f}(\mathbf{i})}\sigma_{\mathbf{g}(\mathbf{i})}$ becomes very close to 0. However, the sensitivity of SSIM rapidly decreases toward zero as C_3 is increased, effectively demonstrating the stabilizing influence of this constant. This is highly desirable as sensitivity near 0 implies that the index does not change significantly as the constant changes.

Denoting the sensitivity of VIF with respect to κ using $P_{\text{VIF}}(\kappa)$, I have

$$
\begin{aligned}
P_{\mathrm{VIF}}(\kappa) &= \frac{\partial \mathrm{VIF}[\mathbf{x}(\mathbf{i},k),\mathbf{y}(\mathbf{i},k)]}{\partial \kappa} && (3.53)\\
&= \frac{\log\left(\frac{\kappa}{\sigma^2_{\mathbf{x}(\mathbf{i},k)}+\kappa}\right)\left[\frac{\hat{\rho}[\mathbf{x}(\mathbf{i},k),\mathbf{y}(\mathbf{i},k)]^2\sigma^2_{\mathbf{y}(\mathbf{i},k)}}{[\sigma^2_{\mathbf{y}(\mathbf{i},k)}(1-\hat{\rho}[\mathbf{x}(\mathbf{i},k),\mathbf{y}(\mathbf{i},k)]^2)+\kappa][\sigma^2_{\mathbf{y}(\mathbf{i},k)}+\kappa]}\right]}{\left[\log\left(\frac{\kappa}{\sigma^2_{\mathbf{x}(\mathbf{i},k)}+\kappa}\right)\right]^2} && (3.54)\\
&- \frac{\log\left(1-\frac{\hat{\rho}[\mathbf{x}(\mathbf{i},k),\mathbf{y}(\mathbf{i},k)]^2\sigma^2_{\mathbf{y}(\mathbf{i},k)}}{\sigma^2_{\mathbf{y}(\mathbf{i},k)}+\kappa}\right)\left[\frac{\sigma^2_{\mathbf{x}(\mathbf{i},k)}}{\kappa(\sigma^2_{\mathbf{x}(\mathbf{i},k)}+\kappa)}\right]}{\left[\log\left(\frac{\kappa}{\sigma^2_{\mathbf{x}(\mathbf{i},k)}+\kappa}\right)\right]^2} && (3.55)
\end{aligned}
$$

The sensitivity of VIF is plotted in Fig. 3.4 as a function of $\sigma^2_{\mathbf{x}(\mathbf{i},k)}$ and κ, with $\sigma^2_{\mathbf{y}(\mathbf{i},k)} = 1$ and $\hat{\rho}[\mathbf{x}(\mathbf{i},k),\mathbf{y}(\mathbf{i},k)] = 0.5$. It can be verified that the sensitivity of VIF goes to 0 as $\sigma^2_{\mathbf{x}(\mathbf{i},k)}$ goes to ∞. However, using L'Hospital's rule, it is seen that the sensitivity of VIF goes to $\pm\infty$ as $\sigma^2_{\mathbf{x}(\mathbf{i},k)}$ goes to 0 even if $\kappa \neq 0$. The sign depends on the values of κ, $\hat{\rho}[\mathbf{x}(\mathbf{i},k),\mathbf{y}(\mathbf{i},k)]$ and $\sigma^2_{\mathbf{y}(\mathbf{i},k)}$. Thus, the VIF index is unstable when the variance of the reference image coefficients is very small, and the constant κ does not assist in stabilizing the index in this region.

It is interesting to note the similarities between the sensitivities of SSIM and VIF. The sensitivity of both indices goes to zero as the variance of the reference image pixels (in SSIM)/coefficients (in VIF) goes to ∞. The sensitivity of both indices to the constant is higher in regions of very low signal energy. Both these observations are qualitatively consistent with my interpretation of the role of the constants, which is to account for masking effects in low signal energy regions.

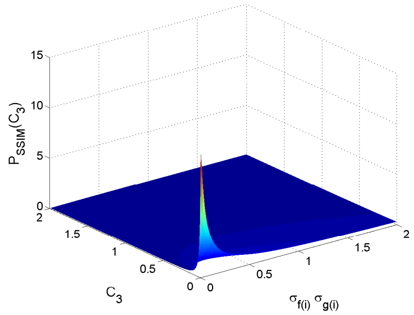


Figure 3.3: Plot of $P_{\text{SSIM}}(C_3)$ as a function of $\sigma_{\mathbf{f}(\mathbf{i})}\sigma_{\mathbf{g}(\mathbf{i})}$ and C_3

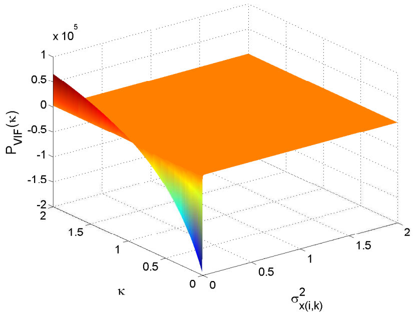


Figure 3.4: Plot of $P_{\text{VIF}}(\kappa)$ as a function of $\sigma^2_{\mathbf{x}(\mathbf{i},k)}$ and κ

3.3.3.2 Vector Model

My interpretation of the normalization performed by the vector VIF model is very similar to that of the scalar model and I discuss it briefly. Once again, the normalization serves two purposes - accounts for masking effects in regions of low signal energy, and incorporates a contrast comparison term in addition to the structure comparison performed by IFC. The instabilities deduced using sensitivity analysis in the scalar VIF index in regions of small reference image energy also occur in the vector VIF model. This is because $Z(\mathbf{i},k)^2\lambda_j(k)$ in (3.31) and (3.33) represents the energy of the j^{th} component of the vector of reference coefficients in a new coordinate system defined by the eigen vectors of $\mathbf{C}_{\mathbf{U}}(k)$.

3.3.4 Discussion

I discussed the relationship between the information theoretic metrics, SSIM metrics and HVS based metrics that have been proposed for IQA. Similarities between the scalar IFC index and HVS based metrics were also observed in [24]. However, my analysis framework and conclusions differ significantly from [24]. In particular, [24] does not discuss the relation between the vector IFC index and HVS based metrics or the relation between the IFC and SSIM indices. Additionally, my discussion of contrast masking and contrast gain control models delves deeper than [24] into the similarities between these HVS based mechanisms and the divisive normalization interpretation of IFC, and leads to very different conclusions discussed below.

I showed that the scalar IFC metric is a monotonic function of the square of the structure term of the SSIM index when the SSIM index is applied on sub-band filtered coefficients. The reasons for the monotonic relationship between the SSIM index and the IFC index are the explicit assumption of a Gaussian distribution on the reference and test image coefficients in the IFC index (conditioned on certain estimated parameters) and the implicit assumption of a Gaussian distribution in the SSIM index (due to the use of regression analysis). With these assumptions in place, the mutual information used as the quality index in the IFC index becomes equivalent to the correlation coefficient used in the SSIM index.

I generalized the concept of the correlation coefficient in SSIM using canonical correlation analysis and established a monotonic relation between the squares of the canonical correlation coefficients and the vector IFC index. Once again, this relation was a direct consequence of two assumptions: a) The reference coefficients are Gaussian distributed when conditioned on the mixing field. b) The use of a linear channel model results in the reference and test image coefficients being jointly Gaussian distributed (conditioned on the mixing field again). One of the contributions of my work is the generalization of the structural similarity philosophy to obtain the probabilistic SSIM index and the vector SSIM index. The scalar probabilistic SSIM index has also been used in optimization with respect to the SSIM index [41–43]. Use of the canonical correlation coefficient as opposed to the simple correlation coefficient has been proposed as an affine invariant measure of quality very recently [69]. My analysis here motivates the canonical correlation coefficient as a natural extension to the SSIM index from a statistical perspective.

I performed an analysis of the sensitivity of the SSIM and VIF index with respect to the constants used in both models. I believe that the SSIM index is superior to the VIF index in terms of its sensitivity to the constant irrespective of the variance of the reference and test images. Additionally, it is easy to intuitively interpret the role of the constant C_3 in SSIM in low signal energy regions, unlike the constant κ in VIF. The instability of VIF in regions of low signal energy is a definite concern and an avenue for possible improvement, e.g., by introducing a stabilizing influence for such regions.

From the discussion on the relation between the structural similarity metrics and HVS based contrast masking models in Section 3.2.4, the relation between information theoretic metrics and contrast masking models are also apparent. The similarities between contrast masking models and natural scene statistical models are not surprising, since HVS modeling and natural scene modeling are considered dual problems [55, 70]. A growing body of work suggests that there is a strong match between the statistics of natural signals and neural processing of these signals [55, 70]. My results show a similar duality within the IQA framework. On the one hand, the GSM model describes the statistics and dependencies in natural image signals that is used in the information theoretic framework [23]. On the other hand, the SSIM index and contrast gain control models in the HVS attempt to eliminate these very same dependencies by modeling divisive normalization mechanisms in neuronal processing of natural signals, whose guiding design principle is hypothesized to be efficient encoding [52]. My interpretation of the duality of natural scene models in IFC and divisive normalization in HVS based metrics differs significantly from [24], which fails to present a cohesive argument for this duality and simply points out certain similarities between the models. Additionally, [24] argues that it is hard to model correlation among coefficients in HVS based metrics. However, my interpretation indicates that this is possible and in fact, divisive normalization models the dual mechanism and eliminates these dependencies between coefficients by including them in the divisive inhibition pool [52].

The assumption that the information theoretic metrics make that individual sub-bands are independent and can be treated separately is not a very good one. The discussion in Section 3.2.5 showed that masking occurs even when the target and the masker have different orientations and the divisive inhibition mechanism needs to account for these effects. Similarly, the coefficients at adjacent orientations and scales of a linear decomposition have dependencies that are accounted for in the GSM model by including all of these coefficients in a GSM vector. Thus, I believe that the information theoretic metrics can be improved upon by removing the restricting assumption of independence across sub-bands, and by the construction of vectors of coefficients that include adjacent orientations and scales, in addition to adjacent spatial locations.

Having discussed the similarities between the structural similarity and the information theoretic frameworks, I will now discuss the differences between them. The structural similarity metrics use a measure of *linear* dependence between the reference and test image pixels, namely the Pearson product moment correlation coefficient. However, the information theoretic metrics use the mutual information, which is a more general measure of correlation that can capture non-linear dependencies between variables. The reason for the monotonic relation between the square of the structure term of the SSIM index applied in the sub-band filtered domain and the IFC index is due to the assumption that the reference and test image coefficients are jointly Gaussian. Although the information theoretic metrics use a better notion of correlation than the structural similarity philosophy, the form of the relationship between the reference and test images might affect visual quality. As an example, if one test image is a deterministic linear function of the reference image, while another test image is a deterministic parabolic function of the reference, the mutual information between the reference and the test image is identical in both cases. However, it is unlikely that the visual quality of both images are identical. I believe that further investigation of suitable models for the distortion channel and the relation between such channel models and visual quality are required to answer this question.

Finally, I observe that due to the fact that the reference and test images are assumed to be jointly Gaussian, both the Pearson product moment correlation coefficient and the mutual information between these variables are a function of the *Signal to Noise Ratio* (SNR). When (X, Y) are jointly Gaussian, they can be described by a linear relation:

$$Y = \beta X + \gamma + N \tag{3.56}$$

where β,γ are constants and N is a Gaussian random variable with variance σ_n^2 that is independent of X. If σ_x^2 is the variance of X, then the correlation coefficient and mutual information between these variables is just a function of SNR defined as: $\beta^2\sigma_x^2/\sigma_n^2$. The SNR determines the probability of correct detection of a signal embedded in noise in signal detection theory. It is not surprising that SNR is a good indicator of visual quality, since QA can be interpreted as the detection of a signal (original image) embedded in a noisy observation (test image). SNR is also widely used in communication systems theory, since the probability of error in communication is a function of the SNR. Thus, the structural similarity and information theoretic metrics differ significantly from the traditional MSE regime, where the measure of quality is the energy of the noise signal. The success of the SSIM and VIF indices show that SNR is a better indicator of visual quality than the energy of the noise signal, an observation that has been used in image QA in the literature [71]. Further, MSE assumes the noise to be additive, while a gain factor β is incorporated in the SSIM (implicitly due to the use of the correlation coefficient) and VIF (explicitly) indices. Thus, the SSIM and VIF indices can be regarded as using improved channel models (attenuation and additive noise), in comparison to MSE (simply additive noise).

3.4 Conclusion

In this chapter, I analyzed two recent philosophies for full reference image QA in a general probabilistic framework - the Structural SIMilarity (SSIM) and the information theoretic paradigms. I explored the relationship between the SSIM index and models of contrast gain control used in human vision based QA systems. I showed that the structure term of the SSIM index is equivalent to certain models of contrast gain control and can hence, account for contrast masking effects in human vision. I also showed that the structure term of SSIM is equivalent to a local computation of mean squared error between reference and test patches, after normalizing the patches appropriately to account for masking. I studied the relationship between information theoretic metrics and the SSIM index. I showed that the IFC computed locally between the coefficients of the reference and test image patches in a sub-band is a monotonic function of the square of the structure term of the SSIM index computed between these patches in the same sub-band. I studied the relationship between the SSIM index and the VIF criterion and revealed certain instabilities in the VIF index. This analysis attempts to unify diverse approaches to the full reference IQA problem derived from different first principles.

Chapter 4

Spatio-temporal Quality Assessment of Natural Videos

Humans can, almost instantaneously, judge the quality of an image or video that they are viewing, using prior knowledge and expectations derived from viewing millions of time-varying images on a daily basis. The right way to assess quality, then, is to ask humans for their opinion of the quality of an image or video, which is known as subjective assessment of image quality. Indeed, subjective judgment of quality must be regarded as the ultimate standard of performance by which image quality assessment (IQA) or video quality assessment (VQA) algorithms are assessed. Subjective quality is measured by asking a human subject to indicate the quality of an image or video that they are viewing on a numerical or qualitative scale. To account for human variability and to assert statistical confidence, multiple subjects are required to view each image/video, and a Mean Opinion Score (MOS) is computed. While subjective methods are the only completely reliable method of VQA, subjective studies are cumbersome and expensive. For example, statistical significance of the MOS must be guaranteed by using sufficiently large sample sizes; subject naivety must be imposed; the dataset of images/videos must be carefully calibrated; and so on [1, 72]. Subjective VQA is impractical for nearly every application other than benchmarking automatic or objective VQA algorithms.

To develop generic VQA algorithms that work across a range of distortion types, full reference algorithms assume the availability of a "perfect" reference video, while each test video is assumed to be a distorted version of this reference.

The discussion in Chapter 2 highlighted the fact that although current full-reference VQA algorithms incorporate features for measuring spatial distortions in video signals, very little effort has been spent on directly measuring temporal distortions or motion artifacts. As described in Chapter 2, several algorithms utilize rudimentary temporal information by differencing adjacent frames or by processing the video using simple temporal filters before feature computation. However, with the exception of [30, 32], existing VQA algorithms do not attempt to directly compute motion information in video signals to predict quality. However, even in [30, 32], motion information is only used to design weights to pool local spatial quality indices.

Yet, motion plays a very important role in the human perception of moving image sequences [73]. Considerable resources in the human visual system (HVS) are devoted to motion perception. The HVS can accurately judge the velocity and direction of moving objects, skills that are essential to survival. Humans are capable of making smooth pursuit eye movements to track moving objects. Visual attention is known to be drawn to movement in the periphery of vision, which makes humans and other organisms aware of approaching danger [74]. Additionally, motion provides important clues about the shape of three dimensional objects and aids in object identification. All these properties of human vision demonstrate the important role that motion plays in perception, and the success of VQA algorithms depends on their ability to model and account for motion perception in the HVS.

While video signals do suffer spatial distortions, they also can be degraded by severe *temporal* artifacts such as ghosting, motion compensation mismatch, jitter, smearing, mosquito noise (amongst numerous other types), as described in detail in Section 4.1. It is imperative that video quality indices account for the deleterious perceptual influence of these artifacts, if objective evaluation of video quality is to accurately predict subjective judgment. Most existing VQA algorithms are able to capture spatial distortions that occur in video sequences (such as those described in Section 4.1.1), but don't do an adequate job in capturing temporal distortions (such as those described in Section 4.1.2).

I seek to address this by developing a general framework for achieving spatio-spectrally localized multiscale evaluation of dynamic video quality. In this framework, both spatial and temporal (and spatio-temporal) aspects of distortion assessment are accounted for. Video quality is evaluated not only in space and time, but also in space-time, by evaluating motion quality along computed motion trajectories.

I develop a general framework for measuring both spatial and temporal video distortions over multi-scales, and along motion trajectories, while accounting for spatial and temporal perceptual masking effects. Using this framework, I develop a full-reference VQA algorithm which I call the MOtion-based Video Integrity Evaluation index, or MOVIE index. MOVIE integrates explicit motion information into the VQA process by tracking perceptually relevant distortions along motion trajectories, thus augmenting the measurement of spatial artifacts in videos. As I show in the sequel, the performance of this approach is highly competitive with the VQA state-of-the-art.

To supply some understanding of the challenging context of VQA, I describe commonly occurring distortions in digital video sequences in Section 4.1. The development of functioning of the MOVIE index is detailed in Section 4.2. I explain the relationship between the MOVIE model and motion perception in biological vision systems in Section 4.3. I also describe the relationship between MOVIE and existing still-image quality indices SSIM and VIF in that section. The performance of MOVIE relative to the state-of-the-art is presented in Section 4.4, using the publicly-available Video Quality Expert Group (VQEG) FR TV Phase 1 database.

4.1 Distortions in Digital Video

In this section, I discuss the kinds of artifacts that are commonly observed in video sequences [75]. I broadly classify these as spatial and temporal distortions. Of course, temporal distortions create spatial artifacts that may be visible in a frozen video frame; and spatial artifacts may change over time. The distinction that I make between spatial and temporal distortions are that temporal distortions arise from the motion of image intensities. This motion may arise from the movement of objects with a scene, or from the motion of the camera, or some combination thereof.

4.1.1 Spatial Artifacts

Examples of commonly occurring spatial artifacts in video include blocking, ringing, mosaic patterns, false contouring, blur and noise. *Blocking effects* result from block based compression techniques used in several DCT based compressions systems such as MPEG-1, MPEG-2, MPEG-4, H.263 and H.264 and appear as periodic discontinuities in each frame of the compressed video at block boundaries. *Ringing distortions* are visible around edges or contours in frames and appear as a rippling effect moving outward from the edge toward the background. Ringing artifacts are visible in non-block based compression systems such as Motion JPEG-2000 as well. *Mosaic Patterns* are visible in block based coding systems and manifest as mismatches between the contents of adjacent blocks as a result of coarse quantization. *False contouring* occurs in smoothly textured regions of a frame containing gradual degradation of pixel values over a given area. Inadequate quantization levels result in step-like gradations having no physical correlate in the reconstructed frame. *Blur* is a loss of high frequency information and detail in video frames. This can occur due to compression, or as a by-product of image acquisition. *Additive Noise* manifests itself as a grainy texture in video frames. Additive noise arises due to video acquisition and by passage of videos through certain communication channels.

4.1.2 Temporal Artifacts

Examples of commonly occurring temporal artifacts in video include motion compensation mismatch, mosquito noise, stationary area fluctuations, ghosting, jerkiness and smearing. *Motion compensation mismatch* occurs due to the assumption that all constituents of a macro-block undergo identical motion, which might not be true. This is most evident around the boundaries of moving objects and appears as the presence of objects and spatial characteristics that are uncorrelated with the depicted scene. *Mosquito effect* is a temporal artifact seen primarily as fluctuations in light levels in smooth regions of the video surrounding high contrast edges or moving objects. *Stationary area fluctuations* closely resemble the mosquito effect in appearance, but are usually visible in textured stationary areas of a scene. *Ghosting* appears as a blurred remnant trailing behind fast moving objects in video sequences. This is a result of deliberate temporal filtering of video sequences in the pre-processing stages to remove additive noise that may be present in the source. *Jerkiness* results from delays during the transmission of video over a network where the receiver does not possess enough buffering ability to cope with the delays. *Smearing* is an artifact associated with the non-instantaneous exposure time of the acquisition device, where light from multiple points of the moving object are integrated into the recording at different instants of time. This appears as a loss of spatial detail in moving objects in a scene.

It is important to observe that temporal artifacts such as motion compensation mismatch, jitter and ghosting alter the movement trajectories of pixels in the video sequence. Artifacts such as mosquito noise and stationary area fluctuations introduce a false perception of movement arising from temporal frequencies created in the test video that were not present in the reference. The perceptual annoyance of these distortions is closely tied to the process of motion perception and motion segmentation that occurs in the human brain while viewing the distorted video.

4.2 Spatio-temporal framework for video quality assessment

In my framework for VQA, separate components for spatial and temporal quality are defined. First, the reference and test videos are decomposed into spatio-temporal bandpass channels using a Gabor filter family. Spatial quality measurement is accomplished by a method loosely inspired by the SSIM index and the information theoretic methods for IQA [15, 24, 37]. Temporal quality is measured using motion information from the reference video sequence. Finally, the spatial and temporal quality scores are pooled to produce an overall video integrity score. These steps are detailed in the following.

4.2.1 Linear Decomposition

Frequency domain approaches are well suited to the study of human perception of video signals and form the backbone of most IQA and VQA systems. Neurons in the visual cortex and the extra-striate cortex are spatial frequency and orientation selective and simple cells in the visual cortex are known to act more or less as linear filters [76–78]. In addition, a large number of neurons in the striate cortex, as well as Area MT which is devoted to movement perception, are known to be directionally selective; i.e., neurons respond best to a stimulus moving in a particular direction. Thus, both spatial characteristics and movement information in a video sequence are captured by a linear spatio-temporal decomposition.

In my framework for VQA, a video sequence is filtered spatio-temporally using a family of band-pass Gabor filters and video integrity is evaluated on the resulting bandpass channels in the spatio-temporal frequency domain. Evidence indicates that the receptive field profiles of simple cells in the mammalian visual cortex are well modeled by Gabor filters [77]. The Gabor filters that I use in the algorithm I develop later are separable in the spatial and temporal coordinates and several studies have shown that neuronal responses in Area V1 are approximately separable [79–81]. The Gabor filters attain the theoretical lower bound on uncertainty in the frequency and spatial variables and thus, visual neurons approximately optimize this uncertainty [77]. In my context, the use of Gabor basis functions guarantees that video features extracted for VQA purposes will be optimally localized.

Further, the responses of several spatio-temporally separable responses can be combined to encode the local speed and direction of motion of the video sequence [82, 83]. Spatio-temporal Gabor filters have been used in several models of the response of motion selective neurons in the visual cortex [82, 84, 85]. In my implementation of the ideas described here, I utilize the algorithm described in [86] that uses the outputs of a Gabor filter family to estimate motion. Thus, the same set of Gabor filtered outputs is used for motion estimation and for quality computation.

A Gabor filter $h(\mathbf{i})$ is the product of a Gaussian window and a complex exponential:

$$h(\mathbf{i}) = \frac{1}{(2\pi)^{\frac{N}{2}} |\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{\mathbf{i}^T \Sigma^{-1} \mathbf{i}}{2}\right) \exp\left(j \mathbf{U_o}^T \mathbf{i}\right) \tag{4.1}$$

where $\mathbf{i} = (x, y, t)$ is a vector denoting a spatio-temporal location in the video sequence. $\mathbf{U_0} = (U_0, V_0, W_0)$ is the center frequency of the Gabor filter and N denotes the dimensionality of these vectors ($N = 3$). Σ is the covariance matrix of the Gaussian component of the Gabor filter. The Fourier transform of the Gabor filter is a Gaussian with covariance matrix Σ^{-1}:

$$H(\mathbf{u}) = \exp\left(-\frac{(\mathbf{u} - \mathbf{U_0})^T \Sigma (\mathbf{u} - \mathbf{U_0})}{2}\right) \tag{4.2}$$

Here, $\mathbf{u} = (u, v, w)$ denotes the spatio-temporal frequency coordinates.

My implementation uses separable Gabor filters that have equal standard deviations along both spatial frequency coordinates and the temporal coordinate. Σ is a diagonal matrix with equal valued entries along the diagonal that I denote as σ^2. My filter design is very similar to the filters used in [86]. However, my filters have narrower bandwidth and are multi-scale as described below.

All the filters in my Gabor filter bank have constant octave bandwidths. I use $P = 3$ scales of filters, with $N = 35$ filters at each scale. Figure 4.1 shows iso-surface contours of the sine phase component of the filters tuned to the finest scale in the resulting filter bank in the frequency domain. The filters at coarser scales would appear as concentric spheres inside the sphere depicted in Fig. 4.1. I used filters with rotational symmetry and the spatial spread of the Gabor filters is the same along all axes. The filters have an octave bandwidth of 0.5 octaves, measured at one standard deviation of the Gabor frequency response. The center frequencies of the finest scale of filters lie on the surface of a sphere in the frequency domain, whose radius is 0.7π radians per sample. Each of these filters has a standard deviation of 2.65 pixels along both spatial coordinates and 2.65 frames along the temporal axis. In my implementation, the Gabor filters were sampled out to a width of three standard deviations; so the support of the kernels at the finest scale are 15 pixels/frames along the spatial/temporal axes. The filters at the coarsest scale lie on the surface of a sphere of radius 0.35π, have a standard deviation of 5.30 pixels (or frames) and a support of 33 pixels (or frames).

Nine filters are tuned to a temporal frequency of 0 radians per sample corresponding to no motion. The spatial orientations of these filters are chosen such that adjacent filters intersect at one standard deviation; hence the spatial orientations of these filters are chosen to be multiples of 20° in the range

$[0°, 180°)$. Seventeen filters are tuned to horizontal or vertical speeds of $s = 1/\sqrt{3}$ pixels per frame and the temporal center frequency of each of these filters is given by $\rho * \frac{s}{\sqrt{s^2+1}}$ radians per sample, where ρ is the radius of the sphere that the filters lie on [86]. Again, the spatial orientations are chosen such that adjacent filters intersect at one standard deviation and the spatial orientations of these filters are multiples of 22° in the range $[0°, 360°)$. The last nine filters are tuned to horizontal or vertical velocities of $\sqrt{3}$ pixels per frame. The spatial orientations of these filters are multiples of 40° in the range $[0°, 360°)$.

Figure 4.2 shows a slice of the sine phase component of the Gabor filters along the plane of zero temporal frequency ($w = 0$) and shows the three scales of filters with constant octave bandwidths. Figure 4.3 shows a slice of the sine phase component of the Gabor filters along the plane of zero vertical spatial frequency. Filters along the three radial lines are tuned to the three different speeds of $(0, \frac{1}{\sqrt{3}}, \sqrt{3})$ pixels per frame.

Finally, a Gaussian filter is included at the center of the Gabor structure to capture the low frequencies in the signal. The standard deviation of the Gaussian filter is chosen such that it intersects the coarsest scale of bandpass filters at one standard deviation.

4.2.2 Spatial MOVIE Index

My approach to capturing spatial distortions in the video of the kind described in Section 4.1.1 is inspired both by the SSIM index and the informa-

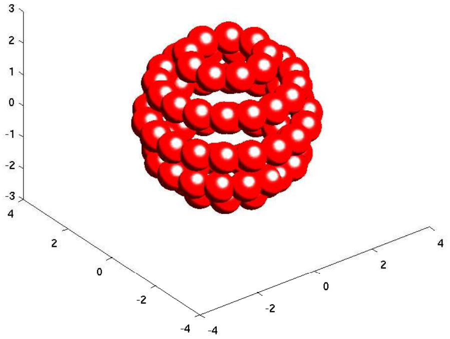


Figure 4.1: Geometry of the Gabor filterbank in the frequency domain. The figure shows iso-surface contours of all Gabor filters at the finest scale. The two horizontal axes denote the spatial frequency coordinates and the vertical axis denotes temporal frequency.

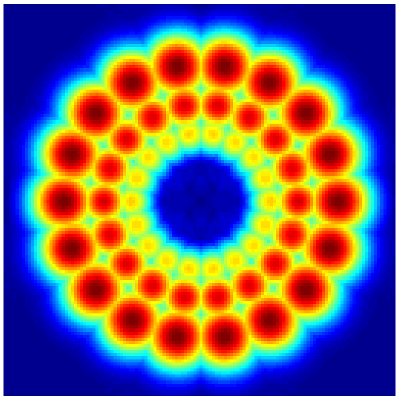


Figure 4.2: A slice of the Gabor filter bank along the plane of zero temporal frequency. The x-axis denotes horizontal spatial frequency and the y-axis denotes vertical spatial frequency.

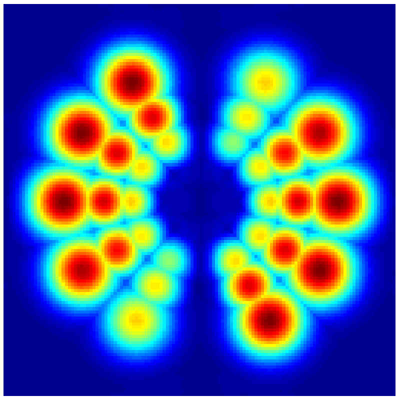


Figure 4.3: A slice of the Gabor filter bank along the plane of zero vertical spatial frequency. The x-axis denotes horizontal spatial frequency and the y-axis denotes temporal frequency.

tion theoretic indices that have been developed for IQA [15, 31, 37]. However, I will be using the outputs of the *spatio-temporal* Gabor filters to accomplish this. Hence, the model described here primarily captures spatial distortions in the video and at the same time, responds to temporal distortions in a limited fashion. I will term this part of my model the "Spatial MOVIE Index", taking this to mean that the model primarily captures spatial distortions. I explain how the Spatial MOVIE index relates to and improves upon prior approaches in Section 4.3.

Let $r(\mathbf{i})$ and $d(\mathbf{i})$ denote the reference and distorted videos respectively. The reference and distorted videos are passed through the Gabor filterbank to obtain band-pass filtered videos. Denote the Gabor filtered reference videos by $\tilde{f}(\mathbf{i}, k)$ and the Gabor filtered distorted videos by $\tilde{g}(\mathbf{i}, k)$, where, $k = 1, 2, \ldots, K$ indexes the filters in the Gabor filterbank. Specifically, let $k = 1, 2, \ldots \frac{K}{P}$

correspond to the finest scale, $k = \frac{K}{P} + 1, \ldots, \frac{2K}{P}$ the second finest scale and so on.

All quality computations begin locally, using local windows B of coefficients extracted from each of the Gabor sub-bands, each spaning N pixels. Consider a pixel location $\mathbf{i}_0$. Let $\mathbf{f}(k)$ be a vector of dimension N, where $\mathbf{f}(k)$ is composed of the *complex magnitude* of N elements of $\tilde{f}(\mathbf{i}, k)$ spanned by the window B centered on $\mathbf{i}_0$. The Gabor coefficients $\tilde{f}(\mathbf{i}, k)$ are complex, but the vectors $\mathbf{f}(k)$ are real and denote the Gabor channel amplitude response [78]. Notice that I have just dropped the dependence on the spatio-temporal location $\mathbf{i}$ for notational convenience by considering a specific location $\mathbf{i}_0$. If the window B is specified by a set of relative indices, then $\mathbf{f}(k) = \{\tilde{f}(\mathbf{i}_0 + \mathbf{m}, k), \mathbf{m} \in B\}$. Similar definition applies for $\mathbf{g}(k)$. To index each element of $\mathbf{f}(k)$, I use the notation $\mathbf{f}(k) = [f_1(k), f_2(k), \ldots, f_N(k)]^T$.

The following spatial quality can then be defined from each subband response:

$$Q_S(\mathbf{i}_0, k) = \frac{1}{2}\frac{1}{N}\sum_{n=1}^{N}\left[\frac{f_n(k) - g_n(k)}{M(k) + C_1}\right]^2 \tag{4.3}$$

where C_1 is a small positive constant added to prevent numerical instability and $M(k)$ is defined as

$$M(k) = \max\left(\sqrt{\frac{1}{N}\sum_{n=1}^{N}|f_n(k)|^2}, \sqrt{\frac{1}{N}\sum_{n=1}^{N}|g_n(k)|^2}\right) \tag{4.4}$$

Notice that the spatial quality is computed as the MSE between $\mathbf{f}(k)$ and $\mathbf{g}(k)$ normalized by a masking function $M(k)$. *Contrast masking* refers to the reduction in visibility of a signal component (target) due to the presence of another signal component of similar frequency and orientation (masker) in a local spatial neighborhood. In the context of VQA, the presence of large signal energy in the image content (masker) masks the visibility of noise or distortions (target) in these regions. The masking function in my model attempts to capture this feature of human visual perception and the masking function is a local energy measure computed from the reference and distorted sub-bands. The outputs of the Gabor filter-bank represent a decomposition of the reference and test video into band-pass channels. Individual Gabor filters respond to a specific range of spatio-temporal frequencies and orientations in the video, and any differences in the spectral content of the reference and distorted videos are captured by the Gabor outputs. Thus, (4.3) will be able to detect primarily spatial distortions in the video such as blur, ringing, false contouring, blocking, noise and so on.

The quality index $Q_S(\mathbf{i}_0, k)$ is bounded and lies between 0 and 1. This follows since

$$\begin{aligned}
Q_S(\mathbf{i}_0, k) &= \frac{1}{2}\frac{1}{N}\sum_{n=1}^{N}\left[\frac{f_n(k) - g_n(k)}{M(k) + C_1}\right]^2 && (4.5)\\
&= \frac{1}{2}\left\{\frac{\frac{1}{N}\sum_{n=1}^{N} f_n(k)^2}{[M(k)+C_1]^2} + \frac{\frac{1}{N}\sum_{n=1}^{N} g_n(k)^2}{[M(k)+C_1]^2} - 2\frac{\frac{1}{N}\sum_{n=1}^{N} f_n(k) g_n(k)}{[M(k)+C_1]^2}\right\} && (4.6)\\
&\leq \frac{1}{2}\left\{\frac{\frac{1}{N}\sum_{n=1}^{N} f_n(k)^2}{[M(k)+C_1]^2} + \frac{\frac{1}{N}\sum_{n=1}^{N} g_n(k)^2}{[M(k)+C_1]^2}\right\} && (4.7)\\
&\leq \left[\frac{M(k)}{M(k)+C_1}\right]^2 && (4.8)
\end{aligned}$$

(4.7) uses the fact that $f_n(k)$ and $g_n(k)$ are non-negative. (4.8) follows from the definition of $M(k)$. Therefore, $Q_S(\mathbf{i}_0, k)$ lies between 0 and 1. Observe that the spatial quality in (4.3) is exactly 0 when the reference and distorted images are identical.

The Gaussian filter responds to the mean intensity or the DC component of the two images. A spatial quality index can be defined using the output of the Gaussian filter operating at DC. Let $\mathbf{f}(\mathrm{DC})$ and $\mathbf{g}(\mathrm{DC})$ denote a vector of dimension N extracted at $\mathbf{i}_0$ from the output of the Gaussian filter operating on the reference and test videos respectively using the same window B. $\mathbf{f}(\mathrm{DC})$ and $\mathbf{g}(\mathrm{DC})$ are low pass filtered versions of the two videos. I first remove the effect of the mean intensity from each video before quality computation, since this acts as a bias to the low frequencies present in the reference and distorted images that are captured by the Gaussian filter. I estimate the mean as the average of the Gaussian filtered output:

$$\mu_{\mathbf{f}} = \frac{1}{N} \sum_{n=1}^{N} f_n(\mathrm{DC}) \tag{4.9}$$

$$\mu_{\mathbf{g}} = \frac{1}{N} \sum_{n=1}^{N} g_n(\mathrm{DC}) \tag{4.10}$$

The quality of the DC sub-band is then computed in a similar fashion as the Gabor sub-bands:

$$Q_{\mathrm{DC}}(\mathbf{i}_0) = \frac{1}{2} \frac{1}{N} \sum_{n=1}^{N} \left[\frac{|f_n(\mathrm{DC}) - \mu_{\mathbf{f}}| - |g_n(\mathrm{DC}) - \mu_{\mathbf{g}}|}{M_{\mathrm{DC}} + C_2} \right]^2 \tag{4.11}$$

where C_2 is a constant added to prevent numerical instability and M_{DC} is defined as

$$M_{\mathrm{DC}} = \max\left(\sqrt{\frac{1}{N}\sum_{n=1}^{N}|f_n(\mathrm{DC}) - \mu_{\mathbf{f}}|^2}, \sqrt{\frac{1}{N}\sum_{n=1}^{N}|g_n(\mathrm{DC}) - \mu_{\mathbf{g}}|^2}\right) \tag{4.12}$$

It is straightforward to verify that $Q_{\mathrm{DC}}(\mathbf{i}_0)$ also lies between 0 and 1. The spatial quality indices computed from all of the Gabor sub-bands and the Gaussian sub-band can then be pooled to obtain a quality index for location $\mathbf{i}_0$ using

$$Q_S(\mathbf{i}_0) = \frac{\sum_{k=1}^{K} Q_S(\mathbf{i}_0, k) + Q_{\mathrm{DC}}(\mathbf{i}_0)}{K+1} \tag{4.13}$$

4.2.3 Motion Estimation

To compute temporal quality, motion information is computed from the reference video sequence in the form of optical flow fields. The same set of Gabor filters used to compute the spatial quality component described above is used to calculate optical flow from the reference video. My implementation uses the successful Fleet and Jepson [86] algorithm that uses the *phase* of the complex Gabor outputs for motion estimation. Notice that I only used the complex magnitude in the spatial quality computation and, as it turns out, I only use the complex magnitudes to evaluate the temporal quality. I have realized a multi-scale version of the Fleet and Jepson algorithm, which I briefly describe in the Appendix.

4.2.4 Temporal MOVIE Index

The spatio-temporal Gabor decompositions of the reference and test video sequences, and the optical flow field computed from the *reference video* using the outputs of the Gabor filters can be used to estimate the temporal video quality. By measuring video quality along motion trajectories, I expect to be able to account for the perceptual effect of distortions of the type described in Section 4.1.2. Once again, the model described here primarily captures temporal distortions in the video, while responding to spatial distortions in a limited fashion. I hence call this stage of my model the "Temporal Movie Index".

First, I discuss how translational motion manifests itself in the frequency domain. Let $a(x, y)$ denote an image patch and let $A(u, v)$ denote its Fourier transform. Assuming that this patch undergoes translation with a velocity $[\lambda, \phi]$ where λ and ϕ denote velocities along the x and y directions respectively, the resulting video sequence is given by $b(x, y, t) = a(x - \lambda t, y - \phi t)$. Then, $B(u, v, w)$, the Fourier transform of $b(x, y, t)$, lies entirely within a plane in the frequency domain [87]. This plane is defined by:

$$\lambda u + \phi v + w = 0 \tag{4.14}$$

Moreover, the magnitudes of the spatial frequencies do not change but are simply sheared:

$$B(u, v, w) = \begin{cases} A(u, v) & \text{if } \lambda u + \phi v + w = 0 \\ 0 & \text{otherwise} \end{cases} \tag{4.15}$$

Spatial frequencies in the video signal provide information about the spatial characteristics of objects in the video sequence such as orientation, texture, sharpness and so on. Translational motion shears these spatial frequencies to create orientation along the temporal frequency dimension without affecting the magnitudes of the spatial frequencies. Translational motion has an easily accessible representation in the frequency domain and these ideas have been used to build motion estimation algorithms for video [82, 83, 87].

Assume that short segments of video without any scene changes consist of local image patches undergoing translation. This is quite reasonable and is commonly used in video encoders that use motion compensation. This model can be used *locally* to describe video sequences, since translation is a linear approximation to more complex types of motion. Under this assumption, the reference and test videos $r(\mathbf{i})$ and $d(\mathbf{i})$ consist of local image patches (such as $a(x, y)$ in the example above) translating to create spatio-temporal video patches (such as $b(x, y, t)$). Observe that (4.14) and (4.15) assume infinite translation of the image patches [87], which is not practical. In actual video sequences, local spectra will not be planes, but will in fact be the convolution of (4.15) with the Fourier transform of a truncation window (a sinc function). However, the rest of my development will assume infinite translation and it will be clear as I proceed that this will not significantly affect the development.

The optical flow computation on the reference sequence provides an estimate of the local orientation of this spectral plane at every pixel of the video. Assume that the motion of each pixel in the distorted video sequence *exactly* matches the motion of the corresponding pixel in the reference. Then, the filters that lie along the motion plane orientation identified from the reference will be activated by the distorted video and the outputs of all Gabor filters that lie away from this spectral plane will be negligible. However, when temporal artifacts are present, the motion in the reference and distorted video sequences do not match. This situation happens, for example, in motion compensation mismatches, where background pixels that are static in the reference move with the objects in the distorted video due to block motion estimation. Another example is ghosting, where static pixels surrounding moving objects move in the distorted video due to temporal low-pass filtering. Other examples are mosquito noise and stationary area fluctuations, where the visual appearance of motion is created from temporal frequencies in the distorted video that were not present in the reference. All of these artifacts shift the spectrum of the distorted video to lie along a different orientation than the reference. Thus, the subset of the Gabor filters that are activated by the distorted video may not be the same as the reference.

Motion vectors from the reference can be used to construct responses from the reference and distorted Gabor outputs that are tuned to the speed and direction of movement of the reference. This is accomplished by computing a weighted sum of the Gabor outputs, where the weight assigned to each individual filter is determined by its distance from the spectral plane of the reference video. Filters that lie very close to the spectral plane are assigned positive excitatory weights. Filters that lie away from the plane are assigned negative inhibitory weights. This achieves two objectives. First, the resulting response is tuned to the movement in the reference video. In other words, a strong response is obtained when the input video has a motion that is equal to the reference video signal. Additionally, any deviation from the reference motion is penalized due to the inhibitory weight assignment. An error computed between these motion tuned responses then serves to evaluate temporal video integrity. The weighting procedure is detailed in the following.

Let $\boldsymbol{\lambda}$ be a vector of dimension N, where $\boldsymbol{\lambda}$ is composed of N elements of the horizontal component of the flow field of the reference sequence spanned by the window B centered on $\mathbf{i}_0$. Similarly, $\boldsymbol{\phi}$ represents the vertical component of flow. Then, using (4.14), the spectrum of the reference video lies along:

$$\lambda_n u + \phi_n v + w = 0, n = 1, 2, \ldots N \tag{4.16}$$

Define a sequence of distance vectors $\boldsymbol{\delta}(k), k = 1, 2, \ldots, K$ of dimension N. Each element of this vector denotes the distance of the center frequency of the k^{th} filter from the plane containing the spectrum of the reference video in a window centered on $\mathbf{i}_0$ extracted using B. Let $\mathbf{U}_0(k) = [u_0(k), v_0(k), w_0(k)], k = 1, 2, \ldots, K$ represent the center frequencies of all the Gabor filters. Then, $\boldsymbol{\delta}(k)$ represents the perpendicular distance of a point from a plane defined by (4.16) in a 3-dimensional space and is given by:

$$\delta_n(k) = \frac{\lambda_n u_0(k) + \phi_n v_0(k) + w_0(k)}{\sqrt{\lambda_n^2 + \phi_n^2 + 1}}, n = 1, 2, \ldots, N \tag{4.17}$$

I now design a set of weights based on these distances. My objective is to assign the filters that intersect the spectral plane to have the maximum weight of all filters. The distance of the center frequencies of these filters from the spectral plane is the minimum of all filters. First, define $\boldsymbol{\alpha}'(k), k = 1, 2, \ldots, K$ using:

$$\alpha'_n k = \frac{\rho(k) - \delta_n(k)}{\rho(k)} \tag{4.18}$$

where $\rho(k)$ denotes the radius of the sphere along which the center frequency of the k^{th} filter lies in the frequency domain.

Figure 4.4 illustrates the geometrical computation specified in (4.18). Each of the circles represents the slice of a Gabor filter in 2 dimensions and the red line shows the projection of the spectral plane in 2 dimensions. The radius $\rho(k)$ and distance $\delta(k)$ are illustrated for one of the Gabor filters.

From the geometry of the Gabor filterbank, it is clear that $0 \leq \alpha'_n(k) \leq 1 \forall n, k$ since the spectral plane specified by (4.16) always passes through the origin. If the spectral plane passes through the center frequency of a Gabor filter k at a location n, then it passes through the corresponding Gabor filter at all scales. $\alpha'_n(k) = 1$ for this filter and the corresponding filters at other scales. If the center frequency of a Gabor filter k at a location n lies along a plane that passes through the origin and is perpendicular to the spectral plane of the reference video, then $\alpha'_n(k) = 0$.

Since I want the weights to be excitatory and inhibitory, I shift all the weights at each scale to be zero-mean [83]. Finally, to make the weights insensitive to the filter geometry that was chosen, I normalize them so that the maximum weight is 1. This ensures that the maximum weight remains 1 irrespective of whether the spectral plane exactly intersects the center frequencies of the Gabor filters. Although the weights are invariant to the filter geometry, observe that due to the Gaussian falloff in the frequency response of the Gabor filters, the Gabor responses themselves are not insensitive to the filter geometry. I hence have a weight vector $\boldsymbol{\alpha}(k), k = 1, 2, \ldots, K$ with elements:

$$\alpha_n(k) = \frac{\alpha'_n(k) - \mu_\alpha}{\max_{k=1,2,\ldots,\frac{K}{P}} \left[\alpha'_n(k) - \mu_\alpha\right]},\ k = 1, 2, \ldots \frac{K}{P} \tag{4.19}$$

where

$$\mu_\alpha = \frac{\sum_{k=1}^{\frac{K}{P}} \alpha'_n(k)}{\frac{K}{P}} \tag{4.20}$$

Similar definitions apply for other scales.

Motion tuned responses from the reference and distorted video sequences may be constructed using these weights. Define N-vectors $\boldsymbol{\nu}^r$ and $\boldsymbol{\nu}^d$ using:

$$\nu_n^r = \frac{|f_n(\mathrm{DC}) - \mu_{\mathbf{f}}|^2 + \sum_{k=1}^{K} \alpha_n(k) f_n(k)^2}{|f_n(\mathrm{DC}) - \mu_{\mathbf{f}}|^2 + \sum_{k=1}^{K} f_n(k)^2 + C_3} \tag{4.21}$$

$$\nu_n^d = \frac{|g_n(\mathrm{DC}) - \mu_{\mathbf{g}}|^2 + \sum_{k=1}^{K} \alpha_n(k) g_n(k)^2}{|g_n(\mathrm{DC}) - \mu_{\mathbf{g}}|^2 + \sum_{k=1}^{K} g_n(k)^2 + C_3} \tag{4.22}$$

The constant C_3 is added to prevent numerical instability.

The vector $\boldsymbol{\nu}^r$ represents the response of the reference video to a mechanism that is tuned to *its own* motion. If the process of motion estimation was perfect and there was infinite translation resulting in a perfect plane, every element of $\boldsymbol{\nu}^r$ would be close to 1. The vector $\boldsymbol{\nu}^d$ represents the response of the distorted video to a mechanism that is tuned to the motion of the *reference video*. Thus, any deviation between the reference and distorted video motions are captured by (4.21) and (4.22).

The denominator terms in (4.21) and (4.22) ensure that temporal quality measurement is relatively insensitive to spatial distortions, thus avoiding redundancy in the spatial and temporal quality measurements. For example, in the case of blur, the same Gabor filters are activated by the reference and distorted videos. However, the response of the finest scale filters are attenuated in the distorted video compared to the reference. Since each video is normalized by its own activity across all filters, the resulting response is not very sensitive to spatial distortions. Instead, the temporal mechanism responds strongly to distortions where there is a misalignment between the spectral planes of the reference and distorted videos.

Finally, the temporal video quality is evaluated by

$$Q_T(\mathbf{i}_0) = \frac{1}{N}\sum_{n=1}^{N}(\nu_n^r - \nu_n^d)^2 \tag{4.23}$$

The temporal quality in (4.23) is also exactly 0 when the reference and test images are identical.

4.2.5 Pooling Strategy

The output of the spatial and temporal quality computation stages is two videos - a spatial quality video $Q_S(\mathbf{i})$ that represents the spatial quality at every pixel of the video sequence and a similar video for temporal quality. The

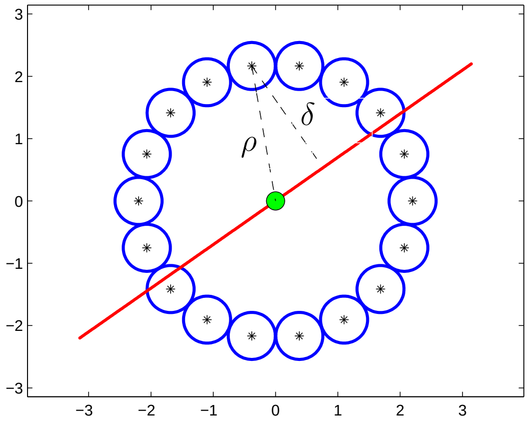


Figure 4.4: A slice of the Gabor filters and the spectral plane shown in 2 dimensions. The horizontal axis denotes horizontal spatial frequencies and the vertical axis denotes temporal frequencies. Each circle represents a Gabor filter and the centers of each filter are also marked. The radius ρ of the single scale of Gabor filters and the the distance δ of the center frequency of one Gabor filter from the spectral plane are marked.

final video quality index, which I call the MOVIE index, combines these into a single VQA index. Consider a set of specific time instants $t = \{t_0, t_1, \ldots, t_\tau\}$ which corresponds to frames in the spatial and temporal quality videos. I refer to these frames of the quality videos, $Q_S(x, y, t_0)$ and $Q_T(x, y, t_0)$ for instance, as “quality maps”.

To obtain a single score for the entire video using the local quality scores obtained at each pixel, several approaches such as probability summation using psychometric functions [7, 10], mean of the quality map [15], weighted summation [30], percentiles [36] and so on have been proposed. In general, the distribution of the quality scores depends on the nature of the scene content and the distortions. For example, distortions tend to occur more in “high activity" areas of the video sequences such as edges, textures and boundaries of moving objects. Similarly, certain distortions such as additive noise affect the entire video, while other distortions such as compression or packet loss in network transmission affect specific regions of the video. Selecting a pooling strategy is not an easy task since the strategy that humans use to determine a quality score based on their perception of an entire video sequence is not known.

I found that the mean of the MOVIE quality maps did not do an adequate job in correlating with visual perception. The quality score assigned to videos that contain a lot of textures, edges, moving objects and so on using the mean of the quality map as the visual quality predictor is consistently worse than quality scores computed for videos that are predominantly smooth. This is because many distortions such as compression alter the appearance of textures and other busy regions of the video much more significantly than the smooth regions of the video. However, people tend to assign poor quality scores even if only parts of the video appear to be distorted.

It would seem apparent that the variance or the spread of the quality scores, in addition to the mean, would prove perceptually significant. Larger variance in the quality scores is indicative of regions of very poor quality in the video, which intuition suggests would result in lower perceptual quality. This is intuitively similar to pooling strategies based on percentiles, wherein the poorest percentile of the quality scores have been used to determine the overall quality [36]. A ratio of the standard deviation to the mean is often used in statistics and is known as the coefficient of variation. The coefficient of variation is a normalized measure of the dispersion of a distribution. Define frame level quality indices for both spatial and temporal components of MOVIE at a frame t_j using:

$$\mathrm{FQ}_S(t_j) = \frac{\sigma_{Q_S(x,y,t_j)}}{1 - \mu_{Q_S(x,y,t_j)}} \tag{4.24}$$

$$\mathrm{FQ}_T(t_j) = \frac{\sigma_{Q_T(x,y,t_j)}}{1 - \mu_{Q_T(x,y,t_j)}} \tag{4.25}$$

The frame level quality indices in (4.24) and (4.25) increases whenever the mean or the standard deviation of the MOVIE scores increases, which is desirable. The use of the standard deviation reduces the content dependent behavior of the mean described earlier. I have found that this moment ratio is a good predictor of the subjective quality of a video.

The spatial MOVIE index is then defined as the average of these frame level descriptors.

$$\text{Spatial MOVIE} = \frac{1}{\tau} \sum_{j=1}^{\tau} \mathrm{FQ}_S(t_j) \tag{4.26}$$

The range of values of the Temporal MOVIE scores is smaller than that of the spatial scores, due to the large divisive normalization in (4.21) and (4.22). To offset this effect, I use the square root of the temporal scores.

$$\text{Temporal MOVIE} = \sqrt{\frac{1}{\tau} \sum_{j=1}^{\tau} \mathrm{FQ}_T(t_j)} \tag{4.27}$$

The MOVIE index is defined as:

$$\text{MOVIE} = \text{Spatial MOVIE} \times \text{Temporal MOVIE} \tag{4.28}$$

4.2.6 Implementation Details and Examples

I now discuss some implementation details of MOVIE. To reduce computation, instead of filtering the entire video sequence with the set of Gabor filters, I centered the Gabor filters on every 16^{th} frame of the video sequence and computed quality maps for only these frames. I selected multiples of 16 since my coarsest scale filters span 33 frames and using multiples of 16 ensures reasonable overlap in the computation along the temporal dimension. The window B was chosen to be a 7×7 window. To avoid blocking artifacts caused by a square window, I used a Gaussian window of standard deviation 1 sampled to a size of 7×7 [15]. If I denote the Gaussian window using $\boldsymbol{\gamma} = \{\gamma_1, \gamma_2, \ldots, \gamma_N\}$ with $\sum_{n=1}^{N} \gamma_n = 1$, (4.3 and (4.4) are modified as:

$$Q_s(\mathbf{i}_0, k) = \frac{1}{2} \sum_{n=1}^{N} \gamma_n \left[\frac{f_n(k) - g_n(k)}{M(k) + C_1} \right]^2 \tag{4.29}$$

$$M(k) = \max \left(\sqrt{\sum_{n=1}^{N} \gamma_n |f_n(k)|^2}, \sqrt{\sum_{n=1}^{N} \gamma_n |g_n(k)|^2} \right) \tag{4.30}$$

Similar modifications apply for (4.9), (4.11) and (4.12). (4.23) is modified as:

$$Q_t(\mathbf{i}_0) = \sum_{n=1}^{N} \gamma_n (\nu_n^r - \nu_n^d)^2 \tag{4.31}$$

There are three parameters in MOVIE: C_1,C_2 and C_3. The role of these constants have been described in detail in [88, 89]. The divisive nature of the masking model in (4.3) and (4.21) makes them extremely sensitive to regions of low signal energy in the video sequences. The constant serves to stabilize the computation in these regions. I selected the constants to be: $C_1 = 0.1$, $C_2 = 1$ and $C_3 = 100$. C_1, C_2 are chosen differently since the Gaussian filter is lowpass and produces larger responses than bandpass Gabor filters. This is intuitively reasonable from the power spectral properties of natural images [90]. C_3 is larger because it is intended to stabilize (4.21) and (4.22), where the denominator terms correspond to *sums* of the squares of all Gabor coefficients. I found that MOVIE is not very sensitive to the choice of constant as long as the constant used is not too small. Using small values for the constants leads to incorrect predictions of poor qualities in smooth regions of the videos due to the instability of the divisive models, which does not match visual perception.

Figures 4.5 and 4.6 illustrate quality maps generated by MOVIE on some representative video sequences. The temporal quality map has been logarithmically compressed for visibility. First of all, it is evident that the kind of distortions captured by the spatial and temporal maps is different. The test video sequences in both examples suffer from significant blurring and the spatial quality map clearly reflects the loss of quality due to blur. The temporal quality map, however, shows poor quality along the edges of objects and in the water where motion compensation mismatches are evident. Of course, the spatial and temporal quality values are not completely independent. This is because the spatial computation uses the outputs of *spatio-temporal* Gabor filters and the constant C_3 in (4.21) and (4.22) permits the temporal computation to respond to blur.

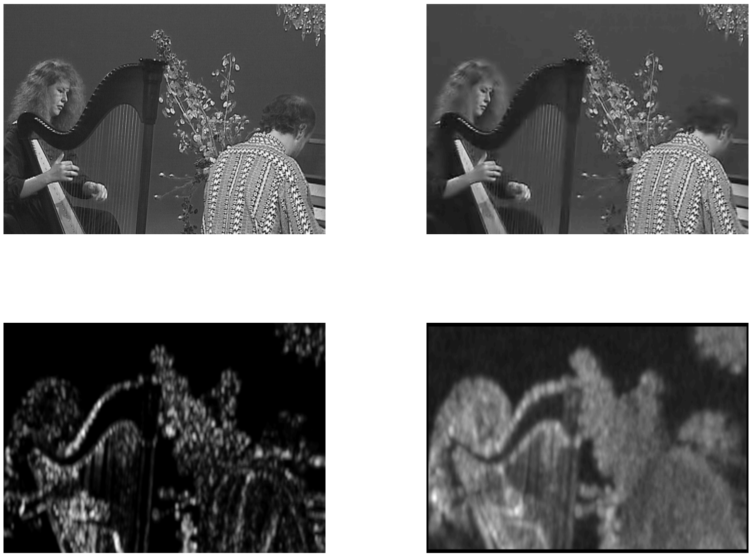


Figure 4.5: Illustration of the performance of the MOVIE index. Top left shows a frame from the reference video. Top right shows the corresponding frame from the distorted video. Bottom left shows a logarithmically compressed temporal quality map. Bottom right shows the spatial quality map. Notice that the spatial quality map responds to the blur in the test video. The temporal quality map responds to motion compensation mismatches surrounding the harp and the heads of the two people and distortions in the strings.

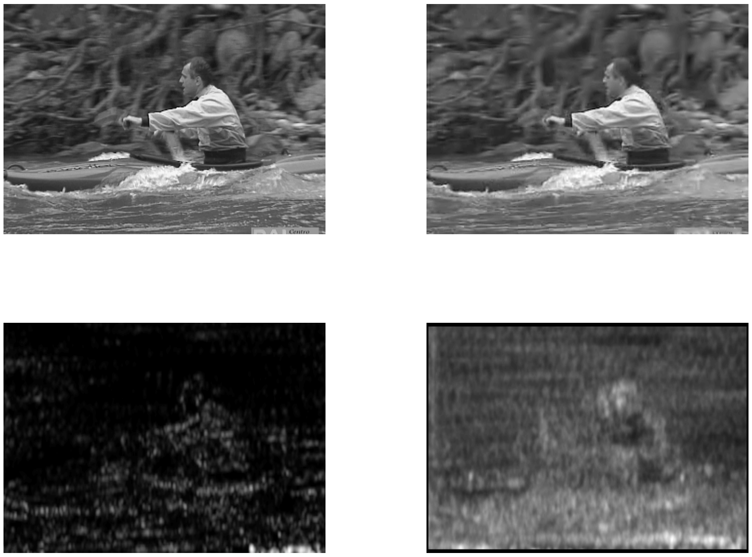


Figure 4.6: Illustration of the performance of the MOVIE index. Top left shows a frame from the reference video. Top right shows the corresponding frame from the distorted video. Bottom left shows a logarithmically compressed temporal quality map. Bottom right shows the spatial quality map. Notice that the spatial quality map responds to the blur in the test video. The temporal quality map responds to motion compensation mismatches surrounding the man, the oar and the ripples in the water.

4.3 Relation to Existing Models

The MOVIE index has some interesting relationships to spatial IQA indices and to visual perception.

4.3.1 Spatial Quality Computation

The spatial quality in (4.3) is closely related to the structure term of the SSIM index [88, 89]. I established the relation between the structure term of the SSIM index and information theoretic methods for IQA in Chapter 3. In particular, I showed that the Gaussian Scale Mixture (GSM) image model assumption used by the information theoretic indices made them equivalent to applying the structure term of the SSIM index in a sub-band domain. Spatial MOVIE falls out naturally from the analysis in Chapter 3 and represents an improved version of these metrics.

I also discussed the relation of both SSIM and IFC to contrast masking models in human vision based IQA systems in Chapter 3. The structure term of the SSIM index applied between sub-band coefficients without the stabilizing constant, assuming zero mean sub-band coefficients, is given by [88, 89]:

$$\frac{1}{2}\frac{1}{N}\sum_{n=1}^{N}\left[\frac{f_n(k)}{\sqrt{\frac{1}{N}\sum_{n=1}^{N}|f_n(k)|^2}}-\frac{g_n(k)}{\sqrt{\frac{1}{N}\sum_{n=1}^{N}|g_n(k)|^2}}\right]^2 \tag{4.32}$$

A chief distinction between SSIM, IFC and the spatial MOVIE index is the fact that I have chosen to utilize both the reference and distorted coefficients to compute the masking term. This is described as ”‘mutual masking”’ in the literature [10]. Masking the reference and test image patches using a measure of their own signal energy in (4.32) ("self masking") is not an effective measure of blur in images and videos. Blur manifests itself as attenuation of certain sub-bands of the reference image and it is easily seen that the self masking model in (4.32) does not adequately capture blur.

However, my model is very different from traditional mutual masking models [10], where the *minimum* of the masking thresholds computed from the reference and distorted images is used. Using a minimum of the masking thresholds is well suited in determining whether the observer can distinguish between the reference and test images, as is done in [10]. However, MOVIE is intended to predict the annoyance of supra-threshold, easily visible distortions. Using the maximum of the two masking thresholds in (4.3) causes the spatial quality index to saturate in the presence of severe distortions (loss of textures, severe blur, severe ringing and so on). This prevents over-prediction of errors in these regions. An additional advantage of using the maximum is that it guarantees bounded quality scores.

4.3.2 Temporal Quality Computation

Motion computation in the HVS is a complex procedure involving low-level and high-level processing. Although motion processing begins in the striate cortex (Area V1), Area MT/V5 in the extra-striate cortex is known to play a significant role in movement processing. The properties of neurons in Area V1 that project to Area MT have been well studied [91]. This study reveals that cells in V1 that project to MT may be regarded as local motion energy filters that are spatio-temporally separable and tuned to a specific frequency and orientation (such as the Gabor filters used here). Area MT receives directional information from V1 and performs more complex computations using the preliminary motion information computed by V1 neurons [91]. A subset of neurons in Area MT have been shown to be *speed tuned*, where the speed tuning of the neuron is independent of the spatial frequency of the stimulus [85, 92]. Models for such speed tuned neurons have been constructed by combining the outputs of a set of V1 cells whose orientation is consistent with the desired velocity [83]. My temporal quality computation bears several similarities with the neuronal model of MT in [83, 93]. Similarities include the weighting procedure based on the distance between the linear filters and the motion plane and the normalization of weighted responses. The models in [83, 93] are rather elaborate, physiologically plausible mechanisms designed to match the properties of visual neurons. My model is designed from an engineering standpoint of capturing distortions in videos. Differences between the two models include the choice of linear decomposition and my derivation of analytic expressions for the weights based on filter geometry. Interestingly, the models of Area MT construct neurons tuned to different speeds and use these responses to determine the speed of the stimulus. My model computes the speed of motion using the Fleet and Jepson algorithm and *then* constructs speed tuned responses based on the computed motion.

To the best of my knowledge, none of the human vision based models for VQA attempt to model properties of neurons in Area MT despite the availability of such models in the vision research community. The discussion here shows that my proposed VQA framework can match visual perception of video better, since it integrates concepts from motion perception.

4.4 Performance

I tested my algorithm on the VQEG FRTV Phase-1 database [2]. Since most of the VQEG videos are interlaced, my algorithm runs on just one field of the interlaced video. I ran my algorithm on the temporally earlier field for all sequences. I ignore the color component of the video sequences, although color might represent a direction for future improvements of MOVIE. The VQEG database contains 20 reference sequences and 16 distorted versions of each reference, for a total of 320 videos. Two distortions types in the VQEG database (HRC 8 and 9) contain 2 different subjective scores assigned by subjects corresponding to whether these sequences were viewed along with "high" or "low" quality videos [2]. I used the scores assigned in the "low" quality regime as the subjective scores for these videos.

The performance of my algorithm is reported for two metrics - the Spearman Rank Order Correlation Coefficient (SROCC) which is an indicator of the prediction monotonicity of the quality index and the Linear Correlation Coefficient (LCC) after non-linear regression. I used the same logistic function specified in [2] to fit the model predictions to the subjective data. The results are reported in Table 4.1.

Prediction Model	SROCC	LCC
Peak Signal to Noise Ratio	0.786	0.779
Proponent P8 (Swisscom)	0.803	0.827
Frame SSIM (no weighting)	0.788	0.820
Frame SSIM (weighted)	0.812	0.849
Spatial MOVIE	0.793	0.796
Temporal MOVIE	0.816	0.801
MOVIE	0.833	0.821

Table 4.1: Comparison of the performance of VQA algorithms.

PSNR provides a baseline for comparison of VQA models. Proponent P8 (Swisscom) is the best performing model of the 10 models tested by the VQEG in terms of both SROCC and LCC after nonlinear regression [2]. Frame SSIM (no weighting) refers to a frame-by-frame application of the SSIM index that was proposed for video in [15]. Frame SSIM (weighted) incorporates rudimentary motion information as weights for different regions of the video sequence [15].

Although my model does not explicitly assume any statistical model for the images or videos, my spatial quality model is closely related to the IFC which assumes that the reference images are the output of a natural scene statistical model [88]. The VQEG database contains 4 sequences that are animated (sources 4,6,16 and 17). Animated videos are quite distinct from natural videos and often contain perfectly smooth and constant regions, perfect step edges, text and so on that seldom occur in natural images. Natural images have several characteristic statistical properties such as self-similarity across scales, heavy tailed wavelet marginal distributions and so on [90, 94],

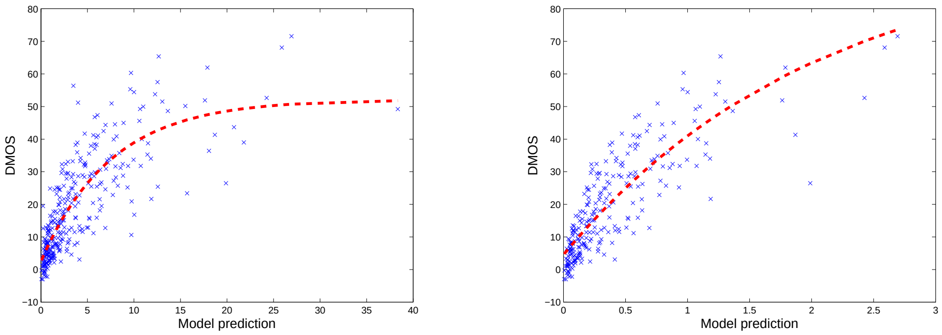
Prediction Model	SROCC	LCC
Spatial MOVIE	0.825	0.830
Temporal MOVIE	0.835	0.825
MOVIE	0.860	0.843

Table 4.2: Performance of MOVIE on the VQEG database after omitting the animation sequences.

that do not occur in synthetic videos of these types. Several aspects of my VQA model such as the choice of Gabor filters, scale invariant processing of the Gabor sub-bands, divisive normalization in the spatial and temporal quality computation are implicitly geared toward natural videos. The presence of text in three of these animations is further cause for concern, since the subjective perception of these videos might have been influenced by the readability of the text in the distorted video. Therefore, I also present performance indices of MOVIE *only* on the 16 natural videos and their distorted versions (a total of 256 videos) in the VQEG database in Table 4.2. I present these results in a separate table since these numbers are not directly comparable against the reported performance of other quality models on all the videos in the database.

Scatter plots of the model prediction and DMOS values, along with the best fitting logistic function, for the MOVIE index are shown in Fig. 4.7.

It is clear that the MOVIE index is competitive with and even outperforms several other systems on the VQEG database. The performance of spatial MOVIE is poorer than that of the temporal MOVIE index, which powerfully illustrates the importance of capturing and assessing temporal video distortions. Using both in conjunction improves over using either separately.



(a) (b)

Figure 4.7: Scatter plot of the subjective DMOS scores against MOVIE scores. The best fitting logistic function used for non-linear regression is also shown. (a) On all sequences in the VQEG database (b) After omitting the animated videos.

The performance of MOVIE is particularly impressive because it does not use any color information and only one field of the interlaced video sequence. All the other models in Table 4.1 use color information.

4.5 Conclusion

The quality of motion representation in videos plays an important role in the perception of video quality, yet existing VQA algorithms make little direct use of motion information, thus limiting their effectiveness. To ameliorate this, I developed a general, spatio-spectrally localized multiscale framework for evaluating dynamic video fidelity that integrates both spatial and temporal (and spatio-temporal) aspects of distortion assessment. Video quality is evaluated not only in space and time, but also in space-time, by evaluating motion quality along computed motion trajectories. Using this framework, I developed a full-reference VQA algorithm known as the MOVIE index. I demonstrated that the MOVIE index delivers VQA scores that correlate quite closely with human subjective judgment, using the VQEG FRTV Phase 1 database as a test bed. Indeed, the MOVIE index was found to be quite competitive with, and even outperform, state-of-the-art VQA algorithms.

Chapter 5

Study of subjective and objective quality assessment of video

The only reliable method to assess the video quality perceived by a human observer is to ask human subjects for their opinion, which is termed subjective quality assessment (QA). Subjective QA is impractical for most applications due to the human involvement in the process. However, subjective QA studies provide valuable data to assess the performance of *objective* or automatic methods of QA. In addition to providing the means to evaluate the performance of state-of-the-art video QA technologies, subjective studies also contribute toward improving the performance of QA methods to reach the ultimate goal of matching human perception.

In this chapter, I first present a study that I conducted to assess the subjective quality of videos. This study included 10 raw naturalistic reference videos and 150 distorted videos obtained from these references using four different real world distortion types. Each video was assessed by 38 human subjects in a single stimulus study, where the subjects scored the quality on a continuous quality scale. This study and the resulting video database presented here, which I call the Laboratory for Image and Video Engineering

(LIVE) Video Quality Database, supplements the popular and widely used LIVE Image Quality Database for still images [95]. I will then present an evaluation of the performance of leading, publicly available objective video QA algorithms on this database.

Currently, to the best of my knowledge, the only publicly available subjective data that is widely used in the video QA community comes from the study conducted by the Video Quality Experts Group (VQEG) as part of it FR-TV Phase I project in 2000 [2]. There have been significant advances in video processing technology since 2000, most notably the development of the H.264/MPEG-4 AVC compression standard that has been adopted widely. The test videos in the VQEG study are not representative of present generation encoders and communication systems. The LIVE video quality database includes videos distorted by H.264 compression, as well as videos resulting from the transmission of H.264 packetized streams through error prone communication channels. Videos obtained from lossy transmission through communication networks exhibit artifacts that are spatially and temporally *transient* and appear as glitches in the video. The LIVE database is unique in this respect, since the VQEG Phase I database does not include such spatio-temporally localized distortion types. Most of the videos in the VQEG study are interlaced, leading to visual artifacts in the reference as well as distorted videos. Objective QA algorithms typically involve multiple processing steps which require adjustment to handle interlaced signals. De-interlacing causes visual artifacts associated with the particular algorithm being used, which is unacceptable in a QA framework. Additionally, interlaced videos are not representative of current trends in the video industry such as multimedia applications, video viewing on computer monitors, progressive High Definition Television (HDTV) standards and so on. The videos in the LIVE Video QA Database are all captured in progressive scan formats, allowing researchers to focus on developing algorithms for QA. Also, the VQEG database was designed to address the needs of secondary distribution of television and hence, the database spans a narrow range of quality scores - more than half of the sequences are of very high quality (MPEG-2 encoded at $>$ 3Mbps). The LIVE database spans a much wider range of quality - the low quality videos in this database were designed to be of similar quality as videos typical of streaming video applications on the Internet (for example, Youtube).

Although the VQEG has several other completed and ongoing projects, the video sequences from none of the subsequent studies have been made public [96]. The videos in my study represent a wide variety of video content and distortion types that are representative of present generation video processing and communication systems. The videos in the LIVE database have been designed specifically with the intent of challenging objective video QA models. I also intend to make the results of the study and the video database publicly available to facilitate comparative evaluation of newer objective models and to advance the state-of-the-art in perceptual quality evaluation systems.

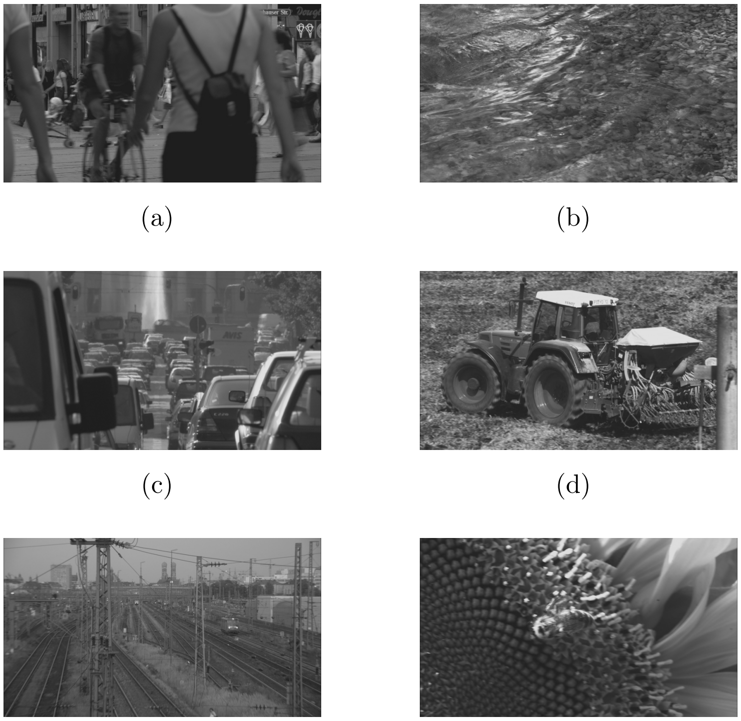
5.1 Details of Subjective Study

5.1.1 Source Sequences

I used ten raw, naturalistic source videos obtained from the Technical University of Munich [97] and the videos are available for download from [98]. All the videos in this database were captured in raw, uncompressed format, which guarantees that the reference videos are distortion free. I only used the progressively scanned videos in this database, to avoid problems associated with video de-interlacing. The videos in the database were captured in High Definition (HD) format. However, due to resource limitations in displaying these videos, I downsampled all videos to a resolution of 768X432 pixels. I chose this resolution to ensure that the aspect ratio of the HD videos are maintained, thus minimizing visual distortions. Additionally, the chosen resolution ensures that the number of rows and columns are multiples of 16, which is often required by compression systems such as MPEG-2. I downsampled each video frame by frame using the "imresize" function in Matlab using bicubic interpolation to minimize distortions due to aliasing.

This database hence contains 10 progressively scanned reference video sequences, all of which have a spatial resolution of 768X432 pixels. Figures 5.1 and 5.2 show one frame of each reference video. All videos, except blue sky, were 10 seconds long. The original blue sky sequence contained only 217 frames and is hence of duration 8.68 seconds. The first seven sequences have a frame rate of 25 frames per second, while the remaining three (Park run, Shields and Mobile & Calendar) have a frame rate of 50 frames per second. A short description of these videos is provided below.

- *Blue Sky* - Circular camera motion showing a blue sky and some trees
- *River Bed* - Still camera, shows a river bed containing some pebbles and water
- *Pedestrian area* - Still camera, shows some people walking about in a street intersection
- *Tractor* - Camera pan, shows a tractor moving across some fields
- *Sunflower* - Still camera, shows a bee moving over a sunflower in close-up
- *Rush hour* - Still camera, shows rush hour traffic on a street
- *Station* - Still camera, shows a railway track, a train and some people walking across the track
- *Park run* - Camera pan, a person running across a park
- *Shields* - Camera pans at first, goes still and then zooms in; shows a person walking across a display pointing at it
- *Mobile & Calendar* - Camera pan, the famous toy train moving across with a calendar moving vertically in the background

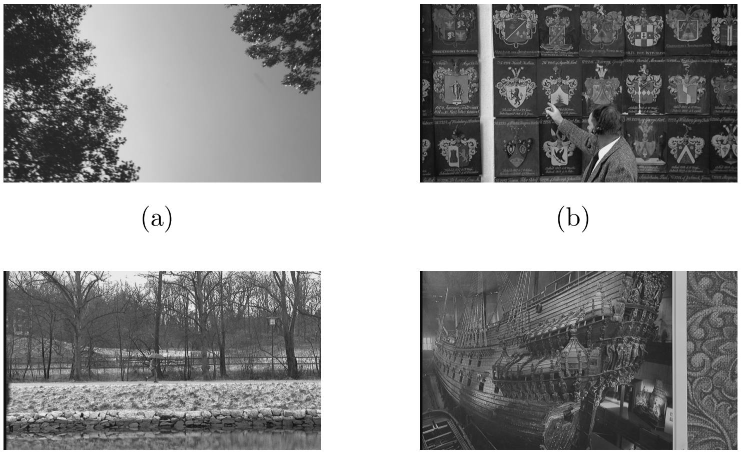


(a) (b)

(c) (d)

(e) (f)

Figure 5.1: One frame from each of the reference video sequences used in the study.



(a) (b) (c) (d)

Figure 5.2: One frame from each of the reference video sequences used in the study.

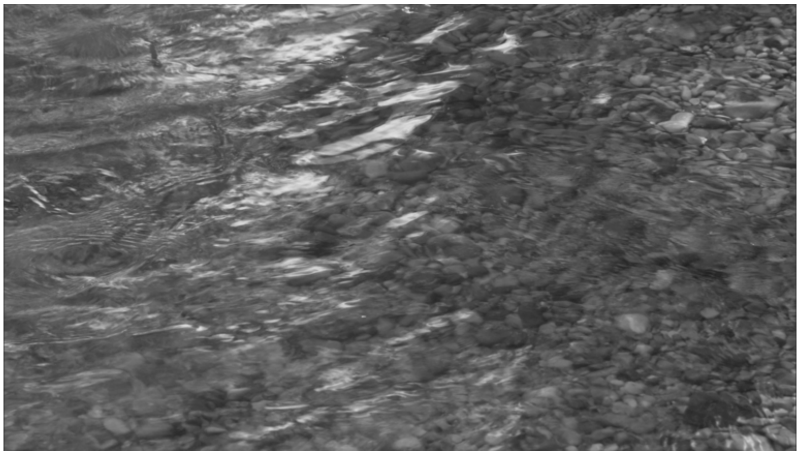
5.1.2 Test Sequences

I created 15 test sequences from each of the reference sequences using four different distortion processes. The goal of this study was to develop a database of videos that will challenge automatic VQA algorithms. I included *diverse* distortion types to test the ability of objective models to predict visual quality consistently across distortions. Compression systems such as MPEG-2 and H.264 produce fairly uniform distortions/quality in the video, both spatially and temporally. Network losses, however, cause *transient* distortions in the video, both spatially and temporally. Figures 5.3,5.4 and 5.5 show one frame of the riverbed sequence from each of the four distortion types in the LIVE database. It is clear that the visual appearance of distortion is very different in each of these videos. MPEG-2 and H.264 compressed videos exhibit typical compression artifacts such as blocking, blur, ringing and motion compensation mismatches around object edges. Videos obtained from lossy transmission through wireless networks exhibit errors that are restricted to small regions of a frame. Videos obtained from lossy transmission through IP networks exhibit errors in larger regions of the frame. Errors in wireless and IP networks are also *temporally transient* and appear as glitches in the video. The LIVE database is unique in this respect, since the VQEG Phase I database does not include such spatio-temporally localized distortion types.

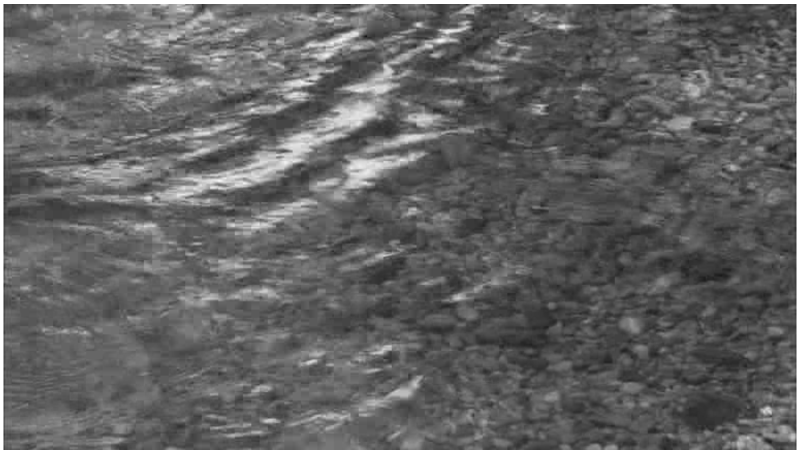
The distortion strengths were adjusted manually, as described in the following, so that the videos obtained from each source and each distortion category spanned the same range of visual quality. This tests the ability of objective VQA models to predict visual quality across content and distortion types consistently. Adjusting distortion strengths perceptually, as I have done here, is far more effective in challenging objective VQA models than fixing the compression rates across sequences as is done in most studies including the VQEG FR-TV Phase I study [2].

5.1.2.1 MPEG-2 compression

The MPEG-2 standard is used in a wide variety of video applications, most notably DVD's and digital television broadcast. I included this distortion type due to its widespread use today. There are four MPEG-2 compressed videos corresponding to each reference in this database. I used the MPEG-2

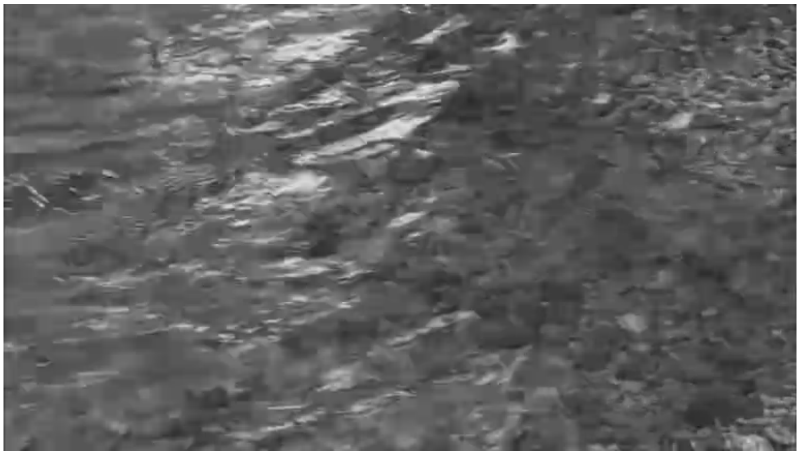


(a)

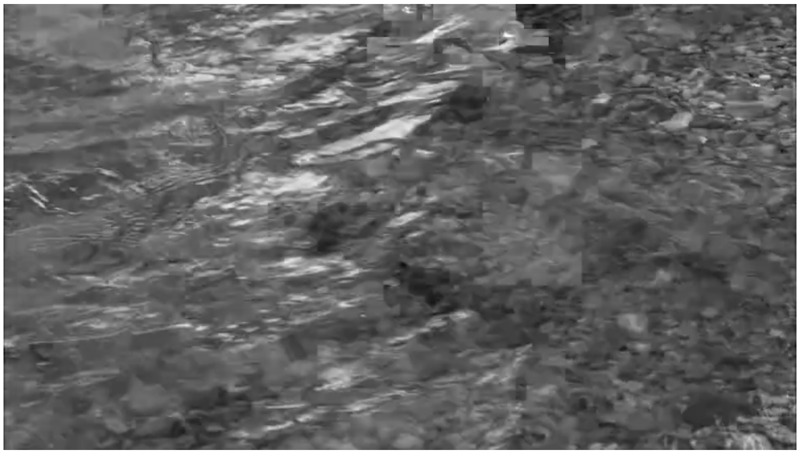


(b)

Figure 5.3: (a) Frame of the reference "riverbed" sequence in the LIVE database (b) Frame from one of the MPEG-2 compressed test video



(a)



(b)

Figure 5.4: (a) Frame from one of the H.264 compressed test video (b) Frame from test video distorted in a simulated IP network

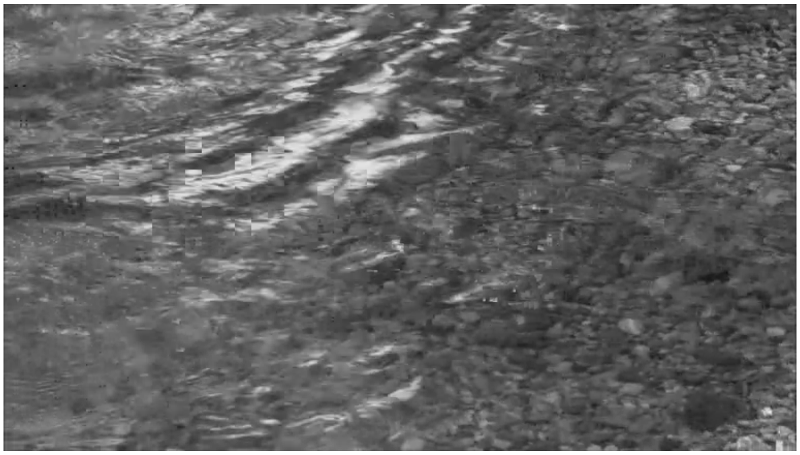


Figure 5.5: Frame from test video distorted in a simulated wireless network

reference software available from the International Organization for Standardization (ISO) to compress the videos [99].

The bit rate required to compress videos for a specified visual quality varies dramatically depending on the content. Compression rates were selected for each reference video such that the test sequences spanned a desired range of perceptual quality. This selection process poses a challenge to objective video QA systems and tests their ability to predict visual quality consistently *across different types of content.*

I viewed compressed videos generated using a wide variety of bit rates to select a suitable subset of four MPEG-2 compressed videos. The “best” video was chosen to be quite close to the reference in visual quality. The “worst” video was chosen to be of poor quality. However, I took care to avoid very low bit rates resulting in videos that are obliterated by MPEG blocking artifacts. The compression rates varied from 700 Kbps to 4 Mbps depending on the reference sequence and further details are provided with the database [100].

5.1.2.2 H.264 compression

H.264 is rapidly gaining popularity due to its superior compression efficiency as compared to MPEG-2. There are four H.264 compressed videos corresponding to each reference in this database. I used the JM reference software (Version 12.3) made available by the Joint Video Team (JVT) [101].

The procedure for selecting the videos was the same as that used to select MPEG-2 compressed videos. Additionally, I ensured that the "best" MPEG-2 and H.264 videos were of similar visual quality and similar considerations applied for the other three pairs. Notice that this test the ability of QA models to consistently predict visual quality *across distortion types* also. The compression rates varied from 200 Kbps to 5 Mbps. I observed the superior performance of H.264 over MPEG-2 and the bit rates used for the "best" and "worst" H.264 videos were lower than their corresponding MPEG-2 counterparts.

5.1.2.3 Transmission over IP Networks

Videos are often transmitted over IP networks in applications such as video telephony and conferencing, IP based streaming and Video on Demand.

There are three “IP” videos corresponding to each reference in my database, that were created by simulating IP losses on an H.264 compressed video stream created using [101].

An in-depth study of the transport of H.264 video over IP networks can be found in [102] and many of my design considerations in the video communication system was based on this study. IP networks offer best effort service and packet losses occur primarily due to buffer overflow at intermediate nodes in a network and congestion. The video sequences subjected to errors in the IP environment contained between one and four slices per frame; I only used these two options since they result in packet sizes that are typical in IP networks. Using one slice per frame has the advantage of reducing overhead due to IP headers at the expense of robustness [102]. Using four slices per frame increases robustness to errors at the expense of reducing compression efficiency.

Four IP error pattern supplied by the Video Coding Experts Group (VCEG), with loss rates of 3%, 5%, 10% and 20%, were used [103]. The error patterns were obtained by real-world experiments and are recommended by the VCEG to simulate the Internet backbone performance for video coding experiments. I created test videos by dropping packets specified in the error pattern from an H.264 compressed packetized video stream. To enable decoding, I did not drop the first packet (containing the Instantaneous Data Refresh (IDR)) and the last packet (since the loss of this packet cannot be detected by the decoder). This is equivalent to assuming that these packets were transmitted reliably out of band. The resulting H.264 bitstream was then decoded using [101] and the losses were concealed using the built-in error concealment mechanism (mode 2 - motion copy) [104].

The procedure for selecting the videos was the same as MPEG-2 compression. However, in this situation, I also paid attention to the type of observed artifact. This is because the observed distortions are different depending on 2 factors

- Whether an Intra-coded frame (I frame) or Predicted frame (P frame) is lost - I frame losses result in much more severe and sustained distortions in the video.
- Whether each frame is transmitted in 1 slice or 4 slices - Loss of an entire frame when transmitted as a single slice results in much more significant distortions, than when the frame is transmitted using 4 slices.

I attempted to pick videos that suffer from different types of observed artifacts, in addition to the consideration that the videos span a desired range of quality as before. Notice that this choice tests the ability of objective QA models to predict the visual annoyance levels across different distortion types, since artifacts arising from compression systems are very different in appearance from artifacts due to network losses.

5.1.2.4 Transmission over wireless networks

Video transmission for mobile terminals is envisioned to be a major application in 3G systems and the superior compression efficiency and error resilience of H.264 makes it ideal for use in harsh wireless transmission environments [105]. There are 4 “wireless” videos corresponding to each reference in the database, that were created by simulating losses sustained by an H.264 compressed video stream (created using [101]) in a wireless environment.

An in-depth study of the transport of H.264 video over wireless networks can be found in [105] and many of my design considerations for the wireless simulations was based on this study. A packet transmitted over a wireless channel is susceptible to bit errors due to attenuation, shadowing, fading and multi-user interference in wireless channels. I assume that a packet is lost even if it contains a single bit error, an assumption that is often made in practice [105]. Due to this assumption, a longer packet is more likely to be lost and I used multiple slices to encode every frame resulting in short packet sizes. I encoded the video stream such that each packet contains roughly the same number of bytes (approximately 200 bytes per packet), making their susceptibility to bit errors almost identical. I simulated errors in wireless environments using bit error patterns and software available from the VCEG [106]. The decoding and error concealment techniques are the same for both wireless and IP simulations.

Observed artifacts in the wireless environment depended on the following factors:

- Whether an I or P frame packet is lost
- Flexible Macroblock Ordering (FMO) - I used both regular and dispersed modes of FMO in my simulations [105]. In dispersed mode, I used four packet groups formed by sub-sampling the frame by 2 along both rows and columns.

Again, I selected videos that suffer from different types of observed artifacts and spanned the desired range of quality. Due to the smaller packet sizes in wireless simulations, the observed artifacts were localized spatio-temporally and appeared different from the artifacts observed in the IP simulations. I used these two different simulation environments to test the ability of objective models to quantify the perceived annoyance of diverse network artifacts.

5.1.3 Subjective Testing Design

I adopted a single stimulus continuous procedure to obtain subjective quality ratings for the different video sequences. The choice of a single stimulus paradigm is more suited for a large number of emerging multimedia applications such as quality monitoring for Video on Demand, internet streaming etc. Additionally, it significantly reduces the amount of time needed to conduct the study for the same number of viewers per sequence, as compared to a double stimulus study. The subjects indicated the quality of the video on a continuous scale. The continuous scale allows the subject to indicate fine gradations in visual quality. I believe this is superior to the ITU-R Absolute Category Scale

(ACR) that uses a 5-category quality judgment which is used in recent VQEG studies [96]. The subject also viewed each of the reference videos to facilitate computation of Difference Mean Opinion Scores (DMOS), a procedure known as hidden reference removal [107, 108].

All the videos in the study were viewed by each subject, which required one hour of the subject's time. To minimize the effects of viewer fatigue, I conducted the study in two sessions of a half hour each, where each subject viewed half the videos.

I prepared playlists for each subject by arranging the 150 test videos in a random order using a random number generator. I did not want the subjects to view successive presentations of test videos that were obtained from the same reference sequence, to avoid contextual and memory effects in their judgment of quality. Once a playlist was constructed, a program would go over the entire playlist to determine if adjacent sequences corresponded to the same content. If any such pairs were detected, one of the videos was swapped with another randomly chosen video in the playlist which did not suffer from the same problem. This list was then split into two halves for the two sessions.

I wanted the subject to view each of the reference videos once in each session for the hidden reference removal process. I inserted each of the ten reference videos to the playlists for each session randomly, again ensuring that successive playback of the same content did not occur.

5.1.4 Subjective Testing Display

I developed the user interface for the study on a Windows PC using MATLAB, in conjunction with the XGL toolbox for MATLAB developed at UT, Austin [109]. The XGL toolbox allows precise presentation of psychophysical stimuli to human observers. It is extremely important to avoid any errors in displaying the video such as latencies or frame drops. This can significantly affect the results of the study since the subject's quality perception is affected not by the video itself, but by the display issues. To ensure perfect playback, all distorted sequences were processed and stored as raw YUV 4:2:0 files. An entire video was loaded into memory before its presentation began to avoid any latencies due to slow hard disk access of large video files. The videos were then played out at the appropriate frame rate for the subject to view. The XGL toolbox interfaces with the ATI Radeon X600 graphics card in the PC and utilizes its ability to play out the YUV videos. The videos were viewed by the subjects on a Cathode Ray Tube (CRT) monitor to avoid the effects of motion blur and low refresh rates on Liquid Crystal Display (LCD) monitors. The entire study was conducted using the same monitor and I calibrated the CRT monitor using the Monaco Optix XR Pro device. The XGL toolbox avoids visual artifacts by synchronizing the display so that the switching between adjacent frames of the video occur during the retrace of the CRT scan. The monitor refresh rate was set at 100 Hz to avoid artifacts due to monitor flicker. Each frame of the 50 Hz videos was displayed for 2 monitor refresh cycles and each frame of the 25 Hz videos was displayed for 4 monitor refresh cycles.

The screen was set at a resolution of 1024×768 pixels and the videos were displayed at their native resolution to prevent any distortions due to scaling operations performed by software or hardware. The remaining areas of the display were black. At the end of the presentation of the video, a continuous scale for video quality was displayed on the screen, with a cursor set at the center of the quality scale to avoid biasing the subject's quality percept. The quality scale had five labels marked on it to help the subject. The left end of the scale was marked "Bad" and the right end was marked "Excellent". Three equally spaced labels between these were marked "Poor", "Fair" and "Good", similar to the ITU-R ACR scale.

The subject could move the cursor along the scale by moving a mouse. The subject was asked to press a key to enter their quality score after moving the cursor to a point on the scale that corresponded to their quality percept. The subject was allowed to take as much time as they needed to enter their score. However, they could not change the score once they had entered it or view the video again. Once the score was entered, the next video was displayed.

5.1.5 Subjects and Training

All subjects taking part in the study were undergraduate students in electrical engineering at the University of Texas at Austin. The subject pool consisted of mostly male students. The subjects were not tested for vision problems. Each video was ranked by 38 subjects.

Each subject was individually briefed about the goal of the experiment. Each subject participated in a short training session at the start of the experiment. The subjects viewed 6 videos in their first session of participation and 3 videos in their second session and were asked to provide quality scores for these videos also to familiarize themselves with the testing procedure. The training videos were not part of the database and contained different content. The training videos were of 10 seconds duration and were also impaired by the same distortions as the test videos. I selected the training videos to span the same range of quality as the test videos.

5.2 Processing of Subjective Scores

Let s_{ijk} denote the score assigned by subject i to video j in session $k = \{1, 2\}$. First, difference scores d_{ijk} are computed by subtracting the quality assigned by the subject to a video from the quality assigned by the same subject to the corresponding reference video in the same session.

$$d_{ijk} = s_{ijk} - s_{ij_{\text{ref}}k} \tag{5.1}$$

The difference scores for the reference videos are 0 in both sessions and are removed from the matrix. The difference scores *per session* are then converted to Z-scores [110]:

$$\mu_{ik} = \frac{1}{N_{ik}} \sum_{j=1}^{N_{ik}} d_{ijk} \tag{5.2}$$

$$\sigma_{ik} = \sqrt{\frac{1}{N_{ik} - 1} \sum_{j=1}^{N_{ik}} (d_{ijk} - \mu_{ik})^2} \tag{5.3}$$

$$z_{ijk} = \frac{d_{ijk} - \mu_{ik}}{\sigma_{ik}} \tag{5.4}$$

where N_{ik} is the number of test videos seen by subject i in session k.

Every subject sees each test video in the database exactly once, either in the first session or in the second session. The Z-scores from both sessions are then combined to create a matrix $\{z_{ij}\}$ of Z-scores, corresponding to the Z-score assigned by subject i to video j. Therefore, $j = \{1, 2, \ldots, N\}$ indexes over $N = 150$ test videos in the LIVE database.

A subject rejection routine is then run on the Z-scores. The procedure specified for subject rejection in the ITU-R BT 500.11 recommendation for the double stimulus impairment scale method is used to discard scores from unreliable subjects [111]. In my study, 9 out of the 38 subjects were rejected at this stage.

Finally, the subjective quality or the Difference Mean Opinion Score (DMOS) for each video is computed as the mean of the Z-scores from the M remaining subjects after subject rejection.

$$\text{DMOS}_j = \frac{1}{M} \sum_{i=1}^{M} z_{ij} \tag{5.5}$$

5.3 Objective QA Models

The performance of several publicly available objective VQA models was evaluated on the LIVE database. One of the problems I faced was the lack of free availability of many VQA algorithms, since many popular VQA algorithms and tools are licensed and sold for profit. These include the Picture Quality Analyzer from Tektronix [112]; the Perceptual Evaluation of Video Quality (PEVQ) from Opticom [113]; the V-Factor from Symmetricom [114]; VQA solutions from SwissQual [115] and Kwill Corporation [116] and several others [117]. My testing was limited to freely available VQA algorithms due to resource limitations.

I tested the following VQA algorithms on the LIVE database.

- *Peak Signal to Noise Ratio (PSNR)* is a simple function of the Mean Squared Error (MSE) between the reference and test videos and provides a baseline for objective VQA model performance.

- *Structural SIMilarity (SSIM)* is a popular method for quality assessment of still images [15, 17]. I used a frame-by-frame implementation of the SSIM index for video. Matlab and Labview implementations of SSIM are available from [118].

- *Multi-scale SSIM* is an extension of the SSIM paradigm, also proposed for still images [18], that has been shown to outperform the SSIM index. I used a frame-by-frame extension of multi-scale SSIM for video. Matlab code for the multi-scale SSIM index for still images was obtained from the authors.

- *Speed SSIM* is the name I give to the VQA model proposed in [32] that uses the SSIM index in conjunction with statistical models of visual speed perception described in [119]. The framework in [32] is shown to improve the performance of both both PSNR and SSIM. I evaluated the performance of this model with the SSIM index since that was shown to be the better performing model in [32]. A software implementation of this index was obtained from the authors.

- *Visual Signal to Noise Ratio (VSNR)* is a quality assessment algorithm proposed for still images [14]. I used a frame-by-frame implementation of VSNR, available from [120].

- *Video Quality Metric (VQM)* is a VQA algorithm developed at the National Telecommunications and Information Administration (NTIA) [36]. Due to its performance in the VQEG Phase II validation tests, the VQM algorithm was adopted by the American National Standards Institute (ANSI) as a national standard, and as international International Telecommunications Union Recommendations (ITU-T J.144 and ITU-R BT.1683, both adopted in 2004). VQM is freely available for download from [121].

- *MOtion-based Video Integrity Evaluation (MOVIE) index* was described in Chapter 4. For computational reasons, the MOVIE index was not computed at every frame of the video sequence. Instead, the Gabor filters were centered on every 8^{th} frame of the video and the MOVIE index was computed only at these frames. Three different versions of the MOVIE index - the Spatial MOVIE index, the Temporal MOVIE index and the MOVIE index - were tested in my study.

5.4 Performance of Objective Models

I tested the performance of all objective models using two metrics - the Spearman Rank Order Correlation Coefficient (SROCC) which measures the monotonicity of the objective model prediction with respect to human scores and the Pearson Linear Correlation Coefficient (LCC) after non-linear regression, which measures the prediction accuracy. I used the logistic function and the procedure outlined in [2] to fit the objective model scores to the DMOS scores.

Table 5.1 shows the performance of all models in terms of the SROCC separately for each distortion type and for the entire LIVE VQA database. Table 5.2 shows the performance of all models in terms of the LCC separately for each distortion type and for the entire LIVE VQA database after nonlinear regression.

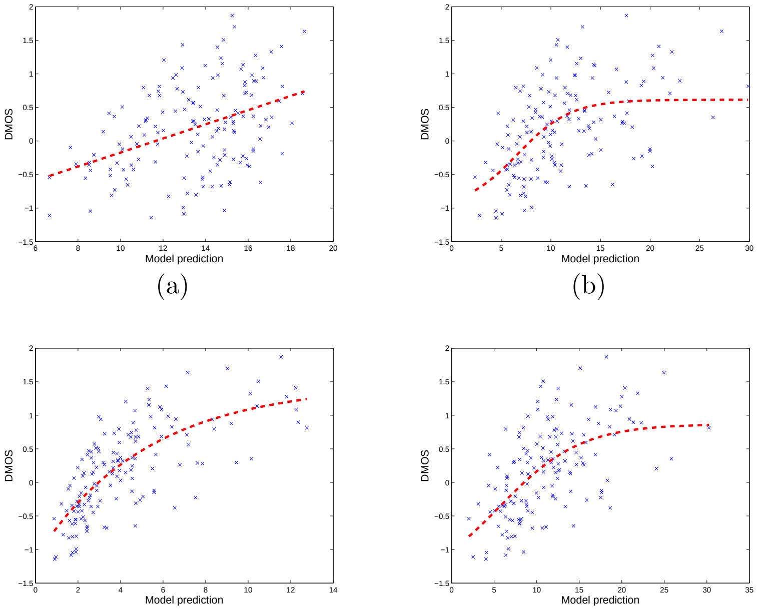
Scatter plots of objective scores vs. DMOS for all the algorithms on the entire LIVE database along with the best fitting logistic functions are shown in Figures 5.6 and 5.7. My results clearly demonstrate that a carefully constructed study can expose the obvious limitations of using PSNR as a VQA

Prediction Model	Wireless	IP	H.264	MPEG-2	All Data
PSNR	0.433	0.321	0.430	0.359	0.368
SSIM	0.522	0.470	0.656	0.561	0.525
Multi-scale SSIM	0.729	0.653	0.705	0.662	0.736
Speed SSIM	0.563	0.473	0.709	0.619	0.585
VSNR	0.702	0.689	0.646	0.591	0.676
VQM	0.721	0.638	0.652	0.781	0.703
Spatial MOVIE	0.793	0.683	0.702	0.697	0.726
Temporal MOVIE	0.647	0.600	0.726	0.804	0.703
MOVIE	0.793	0.671	0.716	0.744	0.740

Table 5.1: Comparison of the performance of VQA algorithms - SROCC

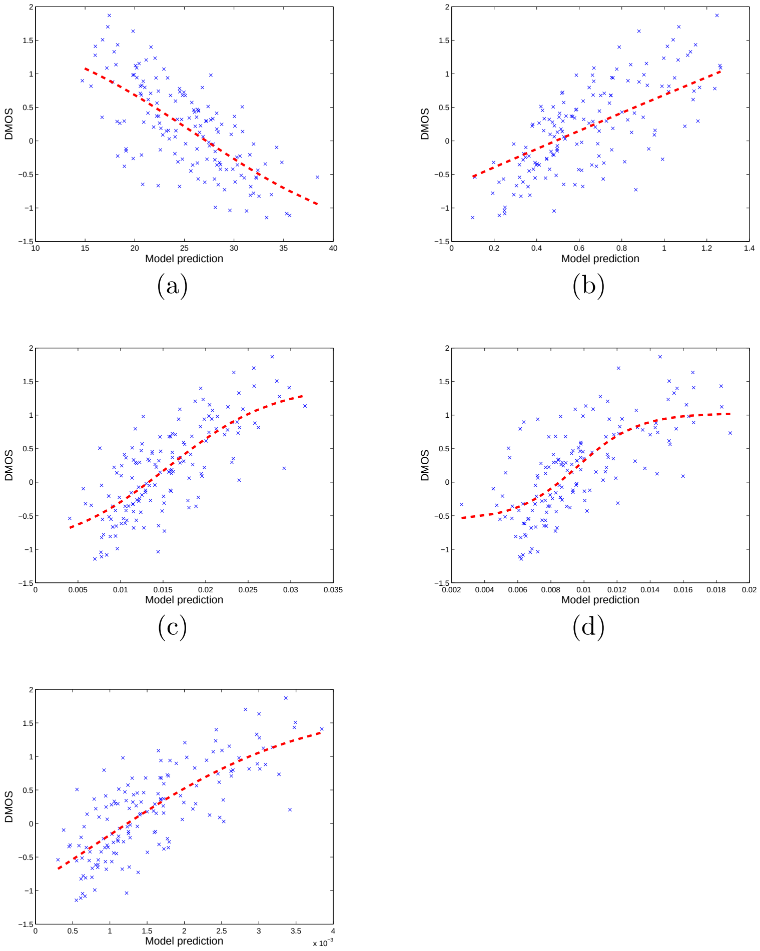
Prediction Model	Wireless	IP	H.264	MPEG-2	All Data
PSNR	0.460	0.411	0.478	0.384	0.401
SSIM	0.546	0.540	0.664	0.576	0.542
Multi-scale SSIM	0.713	0.722	0.688	0.687	0.738
Speed SSIM	0.582	0.578	0.723	0.643	0.596
VSNR	0.699	0.736	0.653	0.675	0.690
VQM	0.735	0.649	0.629	0.797	0.715
Spatial MOVIE	0.795	0.745	0.718	0.714	0.743
Temporal MOVIE	0.659	0.686	0.765	0.817	0.710
MOVIE	0.812	0.730	0.750	0.749	0.761

Table 5.2: Comparison of the performance of VQA algorithms - LCC



(a) (b) (c) (d)

Figure 5.6: Scatter plots of objective VQA scores vs. DMOS for all videos in the LIVE database. Also shown is the best fitting logistic function. (a) PSNR (b) SSIM (c) Multi-scale SSIM (d) Speed SSIM



(a) (b)

(c) (d)

(e)

Figure 5.7: Scatter plots of objective VQA scores vs. DMOS for all videos in the LIVE database. Also shown is the best fitting logistic function. (a) VSNR (b) Spatial MOVIE (c) Temporal MOVIE (d) MOVIE

model. All the VQA models tested in this study improve upon PSNR. Speed SSIM improves upon using just the SSIM index. The best performing VQA algorithm amongst the ones tested in the study are the MOVIE index and the multi-scale SSIM index. VQM emerges as a competitive VQA index, although it did not perform as well as the MOVIE and the multi-scale SSIM indices. VSNR proved to be quite competitive with VQM.

5.5 Conclusions

A subjective study to evaluate the effects of present generation video compression and communication technologies on the perceptual quality of digital video was presented. This study included 150 videos derived from ten reference videos using four distortion types and were evaluated by 38 subjects. The database was unique in terms of content and distortion. I presented an evaluation of the performance of several publicly available objective VQA models on this database.

Chapter 6

Conclusions and Future Work

Successful image quality assessment (IQA) and video quality assessment (VQA) algorithms have the potential to greatly improve digital services for consumers. Monitoring and controlling the quality of broadcast digital video streams is an essential goal for improving the Quality of Service in services such as High Definition TeleVision (HDTV), video on demand, video surveillance, digital cinema, video tele-presence, video phones and other mobile devices.

This dissertation studied objective and subjective methods of perceptual quality evaluation of digital images and video. In Chapter 3, I presented an analysis of the structural and information theoretic methods of IQA and explored their relation to each other, and contrast masking/gain control models in human vision based IQA models. This analysis underscores the similarities between diverse approaches to the IQA problem and provides direction for future research.

In Chapter 4, I proposed a framework for VQA based on motion models. I developed a spatio-temporal, multi-scale framework that can capture spatial, temporal and spatio-temporal distortions in digital video and constructed a

VQA index using this framework known as the MOtion-based Video Integrity Evaluation (MOVIE) index. I also demonstrated the ability of the MOVIE index to predict human opinion scores on a large database of videos.

I studied subjective methods of VQA in Chapter 5. I plan to make the results of my subjective experiments publicly available to the research community to facilitate future research in the development of objective models. The performance of several leading objective VQA algorithms was evaluated using the results of the study. The MOVIE index developed as part of this dissertation was shown to perform very well in this study and shown to be competitive with other state-of-the-art methods.

I will now discuss some directions for future research.

6.1 Pooling Strategies

Two of the distortion types in the database described in Chapter 5 resulting from video transmission through lossy wireless and IP networks cause distortions that are transient, both spatially and temporally. VQA algorithms need to be able to account for such transient distortions. As part of this study, I also recorded quality scores in continuous time provided by the subject as they are viewing the video. This provides a description of the quality of the video as a function of time. This data can be used to design pooling strategies for objective VQA algorithms that can correlate with human data scores.

6.2 Scalable Quality Assessment

The spatial and temporal scale of a video stream is often altered by, e.g., display or transcoding requirements. It is therefore of interest to perform IQA/VQA on videos that have been resolution scaled relative to the reference. Example applications include scalable streaming video over the Internet, video over wireless netowrks, video display on small mobile devices, in-flight entertainment screens, High Definition (HD) videos displayed on Standard Definition (SD) monitors and so on.

When reference and test signals are of different scales, a question that needs to be addressed is whether the test image is to be compared with the original reference (I call this "Scale Adaptive IQA/VQA") or with a modified reference with scale matched to the test image (I call this "Scale Matched IQA/VQA"). In the latter case, the algorithm will assess the annoyance of distortions unrelated to the scaling, while in the former, the rating will also assess quality as a function of resolution loss. Scale Matched IQA/VQA is relevant when the endpoint display device is the limiting factor (cell phones, standard definition televisions etc.). Scale Adaptive IQA/VQA will find broad application owing to the prevalence of scalable video formats that adjust to match the QoS needs of the provider.

Subjective studies of scalable IQA/VQA would be of great interest and value and represent a future direction of research. Further, several full reference IQA techniques discussed in this dissertation - the multi-scale structural similarity index, visual information fidelity index, human vision based models -

perform a decomposition of the reference and test images, often in a multi-scale fashion. The MOVIE framework decomposes the reference and test videos in a multi-scale fashion *spatio-temporally*. Such decompositions facilitate simple solutions for objective scalable IQA/VQA that warrant further investigation.

6.3 Reduced Reference Quality Assessment

Reduce reference IQA/VQA techniques are desirable since availability of a reference video imposes large memory and bandwidth requirements. The MOVIE index can be extended to operate in reduced reference mode, where reference information reduced to different bit rates is used to evaluate the test video quality. One strategy could be to use the motion information computed by MOVIE to select fast moving regions of the video as the reduced reference. This would account for visual tracking of moving objects by the human eye and the increased visibility of distortions in these regions.

6.4 Natural Scene Statistics

Studying probabilistic distributions of images and videos encountered in the natural world on the space of all possible images is an active research area and a large number of applications in image processing and machine vision can benefit from statistical models of the input signals they receive. Natural scene statistical models were used in the information theoretic metrics for still images and I integrated natural image statistics and optical flow to propose a new model for the statistics of video signals in the wavelet domain [122]. The study of the statistical structure of natural scenes can be pursued further and used in applications such as video segmentation and surveillance.

6.5 Blind Quality Assessment

No-reference or blind IQA/VQA algorithms attempt to predict the visual quality of a given image/video without using any other information. This is an extremely challenging problem since it marks a paradigm shift from measuring fidelity to measuring quality. Humans have certain expectations of quality derived from past experience of viewing millions of time-varying imagery using our vision systems. Sophisticated models of natural image and video statistics can be used to develop no reference IQA/VQA algorithms by equating loss of quality with deviation from expected statistics.

Appendix

Appendix 1

Optical Flow Computation Via a New Multi-scale Approach

The Fleet and Jepson algorithm attempts to find constant phase contours of the outputs of a Gabor filterbank to estimate the optical flow vectors [86]. Constant phase contours are computed by estimating the derivative of the phase of the Gabor filter outputs, which in turn can be expressed as a function of the derivative of the Gabor filter outputs [86]. The algorithm in [86] uses a 5-point central difference to perform the derivative computation. However, I chose to perform the derivative computation by convolving the video sequence with filters that are derivatives of the Gabor kernels, denoted by $h'_x(\mathbf{i}), h'_y(\mathbf{i}), h'_t(\mathbf{i})$:

$$h'_x(\mathbf{i}) = h(\mathbf{i}) \left(\frac{-x}{\sigma^2} + jU_0 \right). \tag{1.1}$$

Similar definitions apply for the derivatives along y and t directions. This filter computes the derivative of the Gabor outputs more accurately and produced better optical flow estimates in my experiments.

The original Fleet and Jepson algorithm uses just a single scale of filters. I found that using a single scale of filters was not sufficient, since optical flow was not computed in fast moving regions of the several video sequences due to temporal aliasing [82, 86]. I hence used 3 scales of filters to compute motion by extending the Fleet and Jepson algorithm to multiple scales.

Due to the aperture problem, each Gabor filter is only able to signal the component of motion that is normal to its own spatial orientation. The Fleet and Jepson algorithm computes normal velocity estimates at each pixel for each Gabor filter. Given the normal velocities from the different Gabor outputs, a linear velocity model is fit to each local region using a least squares criterion to obtain a 2D velocity estimate at each pixel of the video sequence. A residual error in the least squares solution is also obtained at this stage. See [86, 123] for further details.

I compute a 2D velocity estimate at each scale using the outputs of the Gabor filters at that scale only. It is important not to combine estimates across scales due to temporal aliasing [82, 86]. I also obtain an estimate of the residual error in the least squares solution for each scale of the Gabor filterbank. The final flow vector at each pixel of the reference video is set to be the 2D velocity computed at the scale with the minimum residual error. Note that more complex solutions such as coarse to fine warping methods have been proposed in the literature to combine flow estimates across scales [124–126]. I chose this approach for simplicity and found that reasonable results were obtained.

The Fleet and Jepson algorithm does not produce flow estimates with 100% density, i.e. flow estimates are not computed at each and every pixel of the video sequence. Instead, optical flow is only computed at pixels where there is sufficient information to do so. I set the optical flow to zero at all pixels where the flow was not computed.

Bibliography

[1] Z. Wang and A. C. Bovik, *Modern Image Quality Assessment.* New York: Morgan and Claypool Publishing Co., 2006.

[2] The Video Quality Experts Group. (2000) Final report from the video quality experts group on the validation of objective quality metrics for video quality assessment. [Online]. Available: http://www.its.bldrdoc.gov/vqeg/projects/frtv_phaseI

[3] K. Seshadrinathan, T. N. Pappas, R. J. Safranek, J. Chen, Z. Wang, H. R. Sheikh, and A. C. Bovik, "Image quality assessment," in *The Essential Guide to Image Processing*, A. C. Bovik, Ed. Elsevier, 2008.

[4] B. Girod, "What's wrong with mean-squared error," in *Digital Images and Human Vision*, A. B. Watson, Ed. The MIT Press, 1993, pp. 207–220.

[5] W. K. Pratt, *Digital Image Processing.* Wiley-InterScience, 1978.

[6] J. Mannos and D. Sakrison, "The effects of a visual fidelity criterion of the encoding of images," *Information Theory, IEEE Transactions on*, vol. 20, no. 4, pp. 525–536, 1974.

[7] J. Lubin, “The use of psychophysical data and models in the analysis of display system performance,” in *Digital Images and Human Vision*, A. B. Watson, Ed. The MIT Press, 1993, pp. 163–178.

[8] P. Burt and E. Adelson, “The laplacian pyramid as a compact image code,” *Communications, IEEE Transactions on*, vol. 31, no. 4, pp. 532–540, 1983.

[9] (2003) JNDMetrix Technology. [Online]. Available: http://www.sarnoff.com/productsservices/videovision/jndmetrix/downloads.asp

[10] S. Daly, “The visible difference predictor: An algorithm for the assessment of image fidelity,” in *Digital Images and Human Vision*, A. B. Watson, Ed. The MIT Press, 1993, pp. 176–206.

[11] A. B. Watson, “The cortex transform: rapid computation of simulated neural images,” *Computer Vision, Graphics, and Image Processing*, vol. 39, no. 3, pp. 311–327, 1987.

[12] P. C. Teo and D. J. Heeger, “Perceptual image distortion,” in *Proceedings of the IEEE International Conference on Image Processing*. IEEE, 1994, pp. 982–986 vol.2.

[13] E. P. Simoncelli, W. T. Freeman, E. H. Adelson, and D. J. Heeger, “Shiftable multiscale transforms,” *Information Theory, IEEE Transactions on*, vol. 38, no. 2, pp. 587–607, 1992.

[14] D. M. Chandler and S. S. Hemami, “VSNR: A wavelet-based visual signal-to-noise ratio for natural images,” *IEEE Trans. Image Process.*, vol. 16, no. 9, pp. 2284–2298, 2007.

[15] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, “Image quality assessment: from error visibility to structural similarity,” *Image Processing, IEEE Transactions on*, vol. 13, no. 4, pp. 600–612, 2004.

[16] L. K. Cormack, “Computational models of early human vision,” in *The handbook of image and video processing*, A. C. Bovik, Ed. New York: Elsevier, 2005, pp. 325–346.

[17] Z. Wang and A. C. Bovik, “A universal image quality index,” *IEEE Signal Process. Lett.*, vol. 9, no. 3, pp. 81–84, 2002.

[18] Z. Wang, E. Simoncelli, A. Bovik, and M. Matthews, “Multiscale structural similarity for image quality assessment,” in *IEEE Asilomar Conference on Signals, Systems and Computers*, 2003, pp. 1398–1402.

[19] Z. Wang and E. P. Simoncelli, “Translation insensitive image similarity in complex wavelet domain,” in *IEEE International Conference onAcoustics, Speech, and Signal Processing*, 2005, pp. 573–576.

[20] R. W. Buccigrossi and E. P. Simoncelli, “Image compression via joint statistical characterization in the wavelet domain,” *Image Processing, IEEE Transactions on*, vol. 8, no. 12, pp. 1688–1701, 1999.

[21] J. Portilla, V. Strela, M. J. Wainwright, and E. P. Simoncelli, “Image denoising using scale mixtures of gaussians in the wavelet domain,” *Image Processing, IEEE Transactions on*, vol. 12, no. 11, pp. 1338–1351, 2003.

[22] A. Srivastava, A. B. Lee, E. P. Simoncelli, and S.-C. Zhu, “On advances in statistical modeling of natural images,” *Journal of Mathematical Imaging and Vision*, vol. 18, no. 1, pp. 17–33, 2003.

[23] M. J. Wainwright and E. P. Simoncelli, “Scale mixtures of gaussians and the statistics of natural images,” in *Adv. Neural Inf. Proc. Sys.*, S. A. Solla, T. Leen, and S.-R. Muller, Eds., vol. 12, 1999.

[24] H. R. Sheikh, A. C. Bovik, and G. de Veciana, “An information fidelity criterion for image quality assessment using natural scene statistics,” *IEEE Trans. Image Process.*, vol. 14, no. 12, pp. 2117–2128, 2005.

[25] C. J. van den Branden Lambrecht and O. Verscheure, “Perceptual quality measure using a spatiotemporal model of the human visual system,” in *Proc. SPIE*, vol. 2668, no. 1. San Jose, CA, USA: SPIE, Mar. 1996, pp. 450–461.

[26] R. E. Fredericksen and R. F. Hess, “Temporal detection in human vision: dependence on stimulus energy,” *Journal of the Optical Society of America A (Optics, Image Science and Vision)*, vol. 14, no. 10, pp. 2557–2569, 1997.

[27] S. Winkler, "Perceptual distortion metric for digital color video," *Proc. SPIE*, vol. 3644, no. 1, pp. 175–184, May 1999.

[28] A. B. Watson, J. Hu, and J. F. McGowan III, "Digital video quality metric based on human vision," *J. Electron. Imaging*, vol. 10, no. 1, pp. 20–29, Jan. 2001.

[29] M. Masry, S. S. Hemami, and Y. Sermadevi, "A scalable wavelet-based video distortion metric and applications," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 16, no. 2, pp. 260–273, 2006.

[30] Z. Wang, L. Lu, and A. C. Bovik, "Video quality assessment based on structural distortion measurement," *Signal Processing: Image Communication*, vol. 19, no. 2, pp. 121–132, Feb. 2004.

[31] H. R. Sheikh and A. C. Bovik, "A visual information fidelity approach to video quality assessment," in *First International conference on video processing and quality metrics for consumer electronics*, 2005.

[32] Z. Wang and Q. Li, "Video quality assessment using a statistical model of human visual speed perception." *J Opt Soc Am A Opt Image Sci Vis*, vol. 24, no. 12, pp. B61–B69, Dec 2007.

[33] M. A. Smith, N. J. Majaj, and J. A. Movshon, "Dynamics of motion signaling by neurons in macaque area MT," *Nat Neurosci*, vol. 8, no. 2, pp. 220–228, Feb. 2005.

[34] J. A. Perrone, "A visual motion sensor based on the properties of V1 and MT neurons," *Vision Research*, vol. 44, no. 15, pp. 1733–1755, Jul. 2004.

[35] The Video Quality Experts Group. (2003) Final VQEG report on the validation of objective models of video quality assessment. [Online]. Available: http://www.its.bldrdoc.gov/vqeg/projects/frtv_phaseII

[36] M. H. Pinson and S. Wolf, "A new standardized method for objectively measuring video quality," *IEEE Transactions on Broadcasting*, vol. 50, no. 3, pp. 312–322, Sep. 2004.

[37] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Trans. Image Process.*, vol. 15, no. 2, pp. 430–444, 2006.

[38] T. N. Pappas and R. J. Safranek, "Perceptual criteria for image quality evaluation," in *Handbook of Image and Video Processing*, A. C. Bovik, Ed. Academic Press, 2005.

[39] S. M. Ross, *Introduction to Probability and Statistics for Engineers and Scientists*. Academic Press, 2004.

[40] D. C. Montgomery and E. A. Peck, *Introduction to Linear Regression Analysis*. John Wiley and sons, 1982.

[41] S. S. Channappayya, A. C. Bovik, and R. W. Heath, Jr., "A linear estimator optimized for the structural similarity index and its application

to image denoising," in *IEEE Intl. Conf. Image Process.*, Atlanta, GA, Jan. 2006.

[42] S. S. Channappayya, A. C. Bovik, C. Caramanis, and R. W. Heath, Jr., "Design of linear equalizers optimized for the structural similarity index," *IEEE Trans. Image Process.*, vol. 17, no. 6, pp. 857–872, 2008.

[43] S. S. Channappayya, A. C. Bovik, and R. W. Heath, Jr., "Rate bounds on SSIM index of quantized images," *IEEE Trans. Image Process.*, to appear, 2008.

[44] S. S. Channappayya, A. C. Bovik, R. W. Heath, Jr., and C. Caramanis, "Rate bounds on the SSIM index of quantized image DCT coefficients," in *Data Compression Conf.*, Snowbird, Utah, Mar. 2008.

[45] J. Nachmias and R. V. Sansbury, "Grating contrast: Discrimination may be better than detection," *Vis. Res.*, vol. 14, no. 10, pp. 1039–1042, Oct. 1974.

[46] G. Legge and J. Foley, "Contrast masking in human vision," *J. Opt. Soc. Am.*, vol. 70, no. 12, pp. 1458–1471, Dec. 1980.

[47] A. Bradley and I. Ohzawa, "A comparison of contrast detection and discrimination," *Vis. Res.*, vol. 26, no. 6, pp. 991–997, 1986.

[48] I. Ohzawa, G. Sclar, and R. D. Freeman, "Contrast gain control in the cat visual cortex," *Nature*, vol. 298, no. 5871, pp. 266–268, Jul. 1982.

[49] D. G. Albrecht and W. S. Geisler, "Motion selectivity and the contrast-response function of simple cells in the visual cortex." *Vis. Neurosci.*, vol. 7, no. 6, pp. 531–546, Dec 1991.

[50] J. Foley, "Human luminance pattern-vision mechanisms: masking experiments require a new model," *J. Opt. Soc. Am. A (Optics and Image Science)*, vol. 11, no. 6, pp. 1710–1719, Jun. 1994.

[51] A. Watson and J. Solomon, "Model of visual contrast gain control and pattern masking," *J. Opt. Soc. Am. A (Optics, Image Science and Vision)*, vol. 14, no. 9, pp. 2379–2391, Sep. 1997.

[52] O. Schwartz and E. P. Simoncelli, "Natural signal statistics and sensory gain control." *Nat. Neurosci.*, vol. 4, no. 8, pp. 819–825, Aug 2001.

[53] R. J. Safranek and J. D. Johnston, "A perceptually tuned sub-band image coder with image dependent quantization and post-quantization data compression," in *IEEE Intl. Conf. Acoustics, Speech and Signal Process.*, Seattle, WA, 1989.

[54] P. Bex and W. Makous, "Spatial frequency, phase, and the contrast of natural images," *J. Opt. Soc. Am. A (Optics, Image Science and Vision)*, vol. 19, no. 6, pp. 1096–1106, Jun. 2002.

[55] V. Mante, R. A. Frazor, V. Bonin, W. S. Geisler, and M. Carandini, "Independence of luminance and contrast in natural scenes and in the

early visual system." *Nat. Neurosci.*, vol. 8, no. 12, pp. 1690–1697, Dec 2005.

[56] V. Bonin, V. Mante, and M. Carandini, "The statistical computation underlying contrast gain control." *J. Neurosci.*, vol. 26, no. 23, pp. 6346–6353, Jun 2006.

[57] A. R. Reibman and D. Poole, "Characterizing packet-loss impairments in compressed video," in *IEEE Intl. Conf. Image Process.*, San Antonio, TX, 2007.

[58] S. de Waele, S. de Waele, and M. J. Verberne, "Coding gain and tuning for parametrized visual quality metrics," in *IEEE Intl. Conf. on Image Process.*, San Antonio, TX, 2007.

[59] J. Ross and H. D. Speed, "Contrast adaptation and contrast masking in human vision," *Proc. Biol. Sci.*, vol. 246, no. 1315, pp. 61–70, Oct. 1991.

[60] R. Shapley and C. Enroth-Cugell, "Visual adaptation and retinal gain controls," *Progress in Retinal Research*, vol. 3, pp. 263–346, 1984.

[61] G. Sclar, J. H. Maunsell, and P. Lennie, "Coding of image contrast in central visual pathways of the macaque monkey," *Vis. Res.*, vol. 30, no. 1, pp. 1–10, 1990.

[62] M. Kivanc Mihcak, I. Kozintsev, K. Ramchandran, and P. Moulin, "Low-complexity image denoising based on statistical modeling of wavelet co-

efficients," *IEEE Signal Processing Letters*, vol. 6, no. 12, pp. 300–303, Dec. 1999.

[63] E. P. Simoncelli, "Modeling the joint statistics of images in the wavelet domain," in *Proc. SPIE*, 1999.

[64] T. M. Cover and J. A. Thomas, *Elements of Information Theory.* John Wiley and sons, 1991.

[65] I. M. Gelfand and A. M. Yaglom, "Calculation of the amount of information about a random function contained in another such function," *Amer. Math. Soc. Transl.*, vol. 12, no. 2, pp. 199–246, 1959.

[66] S. Kullback, *Information Theory and Statistics.* Dover Publications, 1968.

[67] H. Hotelling, "Relations between two sets of variates," *Biometrika*, vol. 28, no. 3-4, pp. 321–377, Dec. 1936.

[68] T. W. Anderson, *An Introduction to Multivariate Statistical Analysis.* John Wiley and Sons, 1984.

[69] H. Takeda, H. J. Seo, and P. Milanfar, "Statistical approaches to quality assessment for image restoration," in *IEEE Intl. Conf. on Consumer Electronics*, Las Vegas, NV, 2008.

[70] E. P. Simoncelli and B. A. Olshausen, "Natural image statistics and neural representation." *Ann. Rev. Neurosci.*, vol. 24, pp. 1193–1216, 2001.

[71] M. Pavel, G. Sperling, T. Riedl, and A. Vanderbeek, “Limits of visual communication: the effect of signal-to-noise ratio on the intelligibility of american sign language,” *J. Opt. Soc. Am. A*, vol. 4, no. 12, pp. 2355–2365, Dec. 1987.

[72] S. Winkler, *Digital Video Quality.* New York: Wiley and Sons, 2005.

[73] B. A. Wandell, *Foundations of Vision.* Sunderland, MA: Sinauer Associates Inc., 1995.

[74] L. Itti, N. Dhavale, and F. Pighin, “Realistic avatar eye and head animation using a neurobiological model of visual attention,” *Proc. SPIE*, vol. 5200, pp. 64–78, 2004.

[75] M. Yuen and H. R. Wu, “A survey of hybrid MC/DPCM/DCT video coding distortions,” *Signal Processing*, vol. 70, no. 3, pp. 247–278, Nov. 1998.

[76] J. A. Movshon, I. D. Thompson, and D. J. Tolhurst, “Spatial summation in the receptive fields of simple cells in the cat’s striate cortex.” *J Physiol*, vol. 283, pp. 53–77, Oct 1978.

[77] J. G. Daugman, “Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters,” *Journal of the Optical Society of America A (Optics and Image Science)*, vol. 2, no. 7, pp. 1160–1169, 1985.

[78] A. C. Bovik, M. Clark, and W. S. Geisler, "Multichannel texture analysis using localized spatial filters," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 12, no. 1, pp. 55–73, Jan. 1990.

[79] D. J. Tolhurst and J. A. Movshon, "Spatial and temporal contrast sensitivity of striate cortical neurones," *Nature*, vol. 257, no. 5528, pp. 674–675, Oct. 1975.

[80] S. M. Friend and C. L. Baker, "Spatio-temporal frequency separability in area 18 neurons of the cat." *Vision Res*, vol. 33, no. 13, pp. 1765–1771, Sep 1993.

[81] M. C. Morrone, M. D. Stefano, and D. C. Burr, "Spatial and temporal properties of neurons of the lateral suprasylvian cortex of the cat." *J Neurophysiol*, vol. 56, no. 4, pp. 969–986, Oct 1986.

[82] D. J. Heeger, "Optical flow using spatiotemporal filters," *International Journal of Computer Vision*, vol. 1, no. 4, pp. 279–302, 1987.

[83] E. P. Simoncelli and D. J. Heeger, "A model of neuronal responses in visual area MT," *Vision Res*, vol. 38, no. 5, pp. 743–761, Mar 1998.

[84] E. H. Adelson and J. R. Bergen, "Spatiotemporal energy models for the perception of motion." *J Opt Soc Am A*, vol. 2, no. 2, pp. 284–299, Feb 1985.

[85] N. J. Priebe, S. G. Lisberger, and J. A. Movshon, "Tuning for spatiotemporal frequency and speed in directionally selective neurons of macaque striate cortex." *J Neurosci*, vol. 26, no. 11, pp. 2941–2950, Mar 2006.

[86] D. Fleet and A. Jepson, "Computation of component image velocity from local phase information," *International Journal of Computer Vision*, vol. 5, no. 1, pp. 77–104, 19900801.

[87] A. B. Watson and J. Ahumada, A. J., "Model of human visual-motion sensing," *Journal of the Optical Society of America A (Optics and Image Science)*, vol. 2, no. 2, pp. 322–342, 1985.

[88] K. Seshadrinathan and A. C. Bovik, "Unifying analysis of full reference image quality assessment," in *IEEE Intl. Conf. on Image Proc.*, accepted for publication.

[89] ——, "Unified treatment of full reference image quality assessment algorithms," *IEEE Trans. Image Process.*, submitted for publication.

[90] D. J. Field, "Relations between the statistics of natural images and the response properties of cortical cells." *J Opt Soc Am A*, vol. 4, no. 12, pp. 2379–2394, Dec 1987.

[91] J. A. Movshon and W. T. Newsome, "Visual response properties of striate cortical neurons projecting to area mt in macaque monkeys," *J. Neurosci.*, vol. 16, no. 23, pp. 7733–7741, 1996.

[92] J. A. Perrone and A. Thiele, “Speed skills: measuring the visual speed analyzing properties of primate mt neurons,” *Nat Neurosci*, vol. 4, no. 5, pp. 526–532, May 2001.

[93] N. C. Rust, V. Mante, E. P. Simoncelli, and J. A. Movshon, “How mt cells analyze the motion of visual patterns.” *Nat Neurosci*, vol. 9, no. 11, pp. 1421–1431, Nov 2006.

[94] E. P. Simoncelli, “Statistical modeling of photographic images,” in *Handbook of Image and Video Processing*, A. C. Bovik, Ed. Academic Press, 2005.

[95] (2003) LIVE image quality assessment database. [Online]. Available: http://live.ece.utexas.edu/research/quality/subjective.htm

[96] (2003) Video Quality Experts Group. [Online]. Available: http://www.its.bldrdoc.gov/vqeg/

[97] Technical university of munich. [Online]. Available: http://www.ei.tum.de/fakultaet/index_html_en

[98] (2003) Video library and tools. [Online]. Available: http://nsl.cs.sfu.ca/wiki/index.php/Video_Library_and_Tools

[99] (2005) International organization for standardization. [Online]. Available: http://standards.iso.org/ittf/PubliclyAvailableStandards/c039486_ISO_IEC_13818-5_2005_Reference_Software.zip

[100] (2008) LIVE video quality assessment database. [Online]. Available: http://live.ece.utexas.edu/research/quality/video.htm

[101] (2007) H.264/AVC software coordination. [Online]. Available: http://iphome.hhi.de/suehring/tml/

[102] S. Wenger, "H.264/AVC over IP," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, no. 7, pp. 645–656, Jul. 2003.

[103] (1999) Proposed error patterns for internet experiments. [Online]. Available: http://ftp3.itu.ch/av-arch/video-site/9910_Red/q15i16.zip

[104] (2007) H.264/MPEG-4 AVC reference software manual. [Online]. Available: http://iphome.hhi.de/suehring/tml/JM%20Reference%20Software%20Manual%20(JVT-X072).pdf

[105] T. Stockhammer, M. M. Hannuksela, and T. Wiegand, "H.264/avc in wireless environments," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, no. 7, pp. 657–673, Jul. 2003.

[106] (1999) Common Test Conditions for RTP/IP over 3GPP/3GPP2. [Online]. Available: http://ftp3.itu.ch/av-arch/video-site/0109_San/VCEG-N80_software.zip

[107] (2008) RRNR-TV Group Test Plan. [Online]. Available: http://www.its.bldrdoc.gov/vqeg/projects/rrnr-tv/

[108] M. H. Pinson and S. Wolf, "Comparing subjective video quality testing methodologies," in *Visual Communications and Image Processing*, T. Ebrahimi and T. Sikora, Eds., vol. 5150, no. 1. SPIE, 2003, pp. 573–582.

[109] (2008) The XGL Toolbox. [Online]. Available: http://128.83.207.86/~jsp/software/xgltoolbox-1.0.5.zip

[110] A. M. van Dijk, J.-B. Martens, and A. B. Watson, "Quality asessment of coded images using numerical category scaling," in *SPIE Advanced Image and Video Communications and Storage Technologies*, vol. 2451, no. 1, 1995.

[111] I.-R. R. BT.500-11, "Methodology for the subjective assessment of the quality of television pictures," International Telecommunications Union, Tech. Rep., 2000.

[112] Tektronix. [Online]. Available: http://www.tek.com/products/video_test/pqa500/

[113] Opticom. [Online]. Available: http://www.opticom.de/technology/pevq_video-quality-testing.html

[114] Symmetricom. [Online]. Available: http://qoe.symmetricom.com/

[115] Swissqual. [Online]. Available: http://www.swissqual.com/Algorithms.aspx

[116] Kwill corporation. [Online]. Available: http://www.kwillcorporation.com/products/VP21H.html

[117] Video quality experts group. [Online]. Available: http://www.its.bldrdoc.gov/vqeg/links/links.php

[118] The structural similarity index. [Online]. Available: http://live.ece.utexas.edu/research/Quality/index.htm

[119] A. A. Stocker and E. P. Simoncelli, "Noise characteristics and prior expectations in human visual speed perception." *Nat Neurosci*, vol. 9, no. 4, pp. 578–585, Apr 2006.

[120] MeTriX MuX Visual Quality Assessment Package. [Online]. Available: http://foulard.ece.cornell.edu/gaubatz/metrix_mux/

[121] VQM. [Online]. Available: http://www.its.bldrdoc.gov/n3/video/VQM_software.php

[122] K. Seshadrinathan and A. C. Bovik, "A structural similarity metric for video based on motion models," in *IEEE Intl. Conf. on Acoustics, Speech, and Signal Proc.*, 2007.

[123] J. L. Barron, D. J. Fleet, and S. S. Beauchemin, "Performance of optical flow techniques," *International Journal of Computer Vision*, vol. 12, no. 1, pp. 43–77, Feb. 1994.

[124] E. P. Simoncelli, "Distributed analysis and representation of visual motion," Ph.D. dissertation, MIT, 1993.

[125] P. Anandan, "A computational framework and an algorithm for the measurement of visual motion," *International Journal of Computer Vision*, vol. 2, no. 3, pp. 283–310, Jan. 1989.

[126] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *Proc. of the 7th Intl. Joint Conf. on Artificial Intelligence*, Vancouver, Canada, 1981, pp. 674–679.

Vita

Kalpana Seshadrinathan was born in Alleppey, India on 29 September 1980. She received the Bachelor of Technology degree in Applied Electronics and Instrumentation Engineering from the University of Kerala, India in 2002. She then joined the Department of Electrical and Computer Engineering at the University of Texas at Austin in Fall 2002 and obtained the Master of Science degree in Summer 2004. She joined the Ph.D. program at the University of Texas at Austin in Fall 2004 under the supervision of Prof. Al Bovik. She joined the Laboratory for Image and Video Engineering (LIVE) as a graduate research assistant in Spring 2003 and was the Assistant Director at LIVE from Spring 2006 until Summer 2008.

Her research interests include image and video quality assessment, statistical modeling of natural images, psychophysics of human vision and motion estimation and representation.

Permanent address: 'DWARAKA', TC. NO. 6/426(7), VARA 131E,
Vattiyoorkavu PO, Thiruvananthapuram,
India - 695013

This dissertation was typeset with LaTeX[†] by the author.

[†] LaTeX is a document preparation system developed by Leslie Lamport as a special version of Donald Knuth's TeX Program.